明　細　書

生物学的物質の配列情報の記録方法及び装置

関連する出願への言及

本願は、２００２年４月１７日に出願された国際出願でＰＣＴ第２１条（２）のもとで英語では公開されていないＰＣＴ／ＪＰ０２／０３８０１の一部継続出願である。本願及びその国際出願は、米国特許法第１１９条のもとで、２００１年４月１８日付け提出の日本国特願２００１－１２０３３５及び２００１年１１月３０日付け提出の日本国特願２００１－３６８００２に対して優先権を主張する。また、明細書、特許請求の範囲、図面、及び要約を含む２００１年４月１８日付け提出の日本国特願２００１－１２０３３５、２００１年１１月３０日付け提出の日本国特願２００１－３６８００２、及び２００２年１０月１６日付け提出の米国特許出願第１０／２７２，１０７号の全ての開示内容は、そっくりそのまま引用して本願に組み込まれている。

技術分野

本発明は、例えばＤＮＡ（デオキシリボ核酸：deoxyribonucleic acid)、ＲＮＡ（リボ核酸：ribonucleic acid）、若しくは遺伝子等の核酸の少なくとも一部を構成する一列のヌクレオチド又はタンパク質の少なくとも一部を構成する一列のアミノ酸などの生物学的物質の配列情報の記録方法及び装置に関する。更に本発明は、その配列情報を供給するためのビジネスモデルとして好適な配列情報の供給方法、及びその配列情報を記録するコンピュータ読み取り可能な記録媒体に関する。

背景技術

人間、及び他の生物（動物、植物、微生物等）のＤＮＡを構成する１対のヌクレオチドの鎖（又は塩基の鎖）の配列情報の解読が世界的に行われている。この場合、従来よりＤＮＡを構成する４種類のヌクレオチドは、塩基としてアデニン

を含むヌクレオチド、グアニンを含むヌクレオチド、シトシンを含むヌクレオチド、及びチミンを含むヌクレオチドにそれぞれ文字Ａ，Ｇ，Ｃ，及びＴを割り当てることによって、それぞれ１バイト（＝８ビット）のテキストデータで表わされている。その結果として一つのＤＮＡの配列は、それを構成する１対の重合体の鎖の内の一方の鎖のヌクレオチド（ｎ個とする）の配列を順次文字Ａ，Ｇ，Ｃ，Ｔ（又はａ，ｇ，ｃ，ｔ）の何れかで表すことによって、ｎバイトのテキストデータで表されていた。同様に、一つのＲＮＡを構成する１本のｎ個のヌクレオチドの配列は、チミンを含むヌクレオチドの代わりにウラシルを含むヌクレオチドに文字Ｕ（又はｕ）を割り当てることによって、ｎバイトのテキストデータで表されていた。

これに関して、例えば人間の最も大きい第１染色体中のＤＮＡの配列は、約２億５千万個のヌクレオチドの配列であり、最も小さい第２２染色体中のＤＮＡの配列は、約５０００万個のヌクレオチドの配列であるため、人間の各染色体中のＤＮＡの配列は、約２５０Ｍバイト～５０Ｍバイトのテキストデータで表すことができる。更に、一人の人間の全部のＤＮＡ情報（ゲノム）は、約３０億個のヌクレオチドの配列で表すことができるため、そのゲノムは、約３Ｇバイトのテキストデータで記録することができる。なお、それらのテキストデータに対して通常のファイル圧縮技術を適用することによって、それらのテキストデータは、例えば元のデータの５０％程度の圧縮ファイルとしても記録、又は送信することができる。

また、ＤＮＡの配列の解読に続いて、ＤＮＡ中の多数の遺伝子の情報に基づいてそれぞれ合成されるタンパク質の機能の研究も広く行われている。この場合、タンパク質を構成する２０種類のアミノ酸は、三文字表記（3-Letter Code）ではそれぞれ３文字（例えばＡｌａ，Ｃｙｓ，Ｇｌｕ等）のテキストデータで表され、一文字表記（1-Letter Code）ではそれぞれ１文字のテキストデータ（例えばＡ，Ｃ，Ｅ等）で表されるため、ｎ個のアミノ酸よりなるタンパク質の配列は、ｎバイトのテキストデータで表すことができる。そして、種々のタンパク質は、それらのアミノ酸が約２０個～約１０００個程度所定の順序で配列されたものであるため、それらのタンパク質の配列は、最大でも約１ｋバイト程度のテキストデー

タで記録することができる。また、例えば人間の遺伝子の総数は約３万個と言われており、タンパク質は理論的なものも含めて約１０万種類の存在が可能であると言われている。

上記の如く例えば一人の人間のＤＮＡ情報をテキストデータで記録するためには、全部で３Ｇバイト程度の記憶容量が必要であり、仮に通常の圧縮ファイルの技術を適用しても１Ｇバイト程度の記憶容量が必要である。また、人間以外の大腸菌や各種ウィルス等のＤＮＡ情報も解析されて次第に公開されるようになっているが、これらのＤＮＡ情報をテキストデータの形で多く集めると、数１００Ｇバイト程度の記憶容量が必要である。これはＲＮＡの配列情報についても同様である。

このように人間又は他の生物のＤＮＡ情報をテキストデータ、又はこの通常の圧縮ファイルの形で記録するものとすると、例えば１枚の記憶容量が５Ｇバイト程度のＤＶＤ－ＲＯＭ(digital video disc-ROM)ディスクのように膨大な記憶容量を持つ記録媒体が必要である。更に、そのＤＮＡ情報を利用する場合にその記録媒体からの読み出し時間が長くなり、処理時間が長くなるという不都合がある。

また、現状の一般の通信回線の通信速度は、最大で５Ｍｂｐｓ程度であるため、例えば１Ｇバイト程度のＤＮＡ情報をその通信回線を介して送信するものとすると、送信時間は最短でも約３０分程度となる。特に最近はそのＤＮＡ情報をデジタルの携帯電話システムを介して送信する場合も考えられるが、現在の携帯電話システムの通信速度はせいぜい１Ｍｂｐｓ程度であるため、少なくとも人間のＤＮＡ情報の伝送で使用することは現状ではあまり実用的ではない。

次に、例えば或る微生物のＤＮＡ中の遺伝子について複数の研究者が並行して研究するような場合に、複数の研究者が保有している標準となるＤＮＡのヌクレオチドの配列の同一性をどのように保証するのかという問題がある。即ち、そのＤＮＡのヌクレオチドの配列が例えば数Ｍバイト（文字数で数１００万文字）程度のテキストデータで記録されている場合に、複数の研究者が互いに自分のテキストデータと他人のテキストデータとの同一性（完全一致性）を短時間に確認するのは必ずしも容易ではない。

これに関連して、例えば人間又は他の生物のＤＮＡ情報の利用方法としては、

標準的なＤＮＡの配列と、検査対象のＤＮＡの配列との間の相違する部分をサーチする場合が考えられる。これは、いわゆるＳＮＰ（一塩基変位多型：Single Nucleotide Polymorphism）の可能性を検査するような場合に必要になると考えられる。しかしながら、両方のＤＮＡのヌクレオチドの配列がそれぞれ膨大なテキストデータで表わされている場合に、それら２つのテキストデータを比較して相違点を検出するにはかなりの長い時間が必要となり、検査時間が長くなるという不都合がある。

更に、人間又は他の生物のＤＮＡ情報を製薬会社の研究者等のユーザに提供するビジネスも行われつつあるが、この場合に、情報供給者が例えば通信回線を介してＤＮＡ情報をユーザに提供する場合には、できるだけ少ない情報量で、即ち短い送信時間で必要な情報をユーザに提供できるビジネスモデルが必要である。また、ユーザ側では、提供されたＤＮＡ情報に伝送エラー等が無いかどうかを容易に確認できることが望ましい。上記の各課題はＲＮＡのヌクレオチドの配列情報についても同様に当てはまるものである。

更に、一つのタンパク質のアミノ酸の配列は、最大でも約１ｋバイト程度のテキストデータで記録することができるが、タンパク質の種類は理論的に約１０万個程度にもなるため、全部のタンパク質の配列情報をテキストデータで表すと、全部のＤＮＡの配列情報程度の膨大な量となる。従って、個々のタンパク質の配列は、できるだけ少ない情報量で記録できることが望ましい。また、２つのタンパク質の配列情報の相違部を容易に確認できるシステムも必要である。

本発明は斯かる点に鑑み、核酸中の一列のヌクレオチド、又はタンパク質中の一列のアミノ酸などの生物学的物質の配列情報を近似的に少ないデータ量で記録できる記録方法及び装置を提供することを第１の目的とする。

また、本発明は、２つの生物学的物質の配列情報同士の相違する部分を少ないデータ量で容易に検出できると共に、必要に応じてその相違する部分の情報を復元できる記録方法及び装置を提供することを第２の目的とする。

また、本発明は、一列のヌクレオチド、又は一列のアミノ酸などの生物学的物質の配列情報をユーザに提供する場合に、ユーザが提供された配列情報と情報供給者が保持している配列情報との相違する部分を少ないデータ量で容易に確認で

きるビジネスモデル（情報供給方法）を提供することを第３の目的とする。

また、本発明は、生物学的物質の配列情報が少ないデータ量で近似的に記録されたコンピュータ読み取り可能な記録媒体を提供することを第４の目的とする。

発明の開示

本発明による第１の生物学的物質の配列情報の記録方法は、その生物学的物質の配列に対応するテキストデータ、又はこのテキストデータを所定の規則に従って変換して得られる数値データを、その生物学的物質の配列方向に複数行で、かつその配列方向に交差する非配列方向に複数列の長さがｍビット（ｍは１６以上の整数）の部分データ（Ａ（ｉ，ｊ））に分割し、複数行のその部分データに各行毎にその非配列方向にガロア体ＧＦ（2^m）上の第１の演算を施して第１組のパリティ情報（Ｂ１（ｉ），Ｂ２（ｉ），Ｂ３（ｉ））を求めると共に、複数列のその部分データに各列毎にその配列方向にガロア体ＧＦ（2^m）上の第２の演算を施して第２組のパリティ情報（Ｃ１（ｊ），Ｃ２（ｊ），Ｃ３（ｊ））を求め、その第１組及び第２組のパリティ情報でその生物学的物質の配列を表すものである。

斯かる本発明によれば、その生物学的物質としては、例えば一列のヌクレオチド又は一列のアミノ酸が考えられる。前者の一列のヌクレオチドは、例えば或るＤＮＡ（deoxyribonucleic acid）を構成する１対の重合体の鎖の一方の鎖の少なくとも一部、或るＲＮＡ（ribonucleic acid）を構成する１列の重合体の鎖の少なくとも一部、又は或る遺伝子の少なくとも一部である。その一列のヌクレオチドの配列は、各ヌクレオチドに含まれる塩基の配列ともみなすことができる。一方、後者の一列のアミノ酸は、例えば或るタンパク質を構成するアミノ酸の配列の少なくとも一部である。

その生物学的物質の全体の個数をＮＴとして、各生物学的物質をそれぞれ１文字（例えばアルファベット）で表すものとすると、その生物学的物質の配列に対応するテキストデータの全体の量は、例えばアスキーコード（ASCII code）（ＡＮＳＩ形式）ではＮＴバイトになり、ユニコード（Unicode）では２・ＮＴバイトになる。なお、配列を見易くするためのスペース、数字、及び改行などのコード

は無視している。そして、例えば図７の例において、テキストデータを配列方向にＮ個（ｉ＝１～Ｎ）で、非配列方向にＭ個（ｊ＝１～Ｍ）の部分テキストデータＴ（ｉ，ｊ）に分割し、図８に示すように、各部分テキストデータＴ（ｉ，ｊ）をそれぞれｍビットの部分データＡ（ｉ，ｊ）に変換する。ｍビットの部分データＡ（ｉ，ｊ）は、それぞれｎ個（図８の例ではｎ＝１６）の連続する生物学的物質の配列を表している。

この場合、最も簡単な方法としては、部分データＡ（ｉ，ｊ）として部分テキストデータＴ（ｉ，ｊ）そのものを数値データとみなしたデータを使用すればよい。即ち、テキストデータがアスキーコードで記録されている場合には、部分データＡ（ｉ，ｊ）としてはそのアスキーコードを使用すればよい。また、テキストデータがユニコードで記録されている場合には、各文字をそれぞれの２バイトのコードの上位１バイトで表したものを部分データＡ（ｉ，ｊ）としてもよい。但し、処理対象のデータ量を少なくするためには、各生物学的物質を表す文字を例えば６ビット以下の数値データに変換する変換テーブル（所定の規則）を用いて、部分テキストデータＴ（ｉ，ｊ）を変換して得られる数値データを部分データＡ（ｉ，ｊ）とすることが望ましい。

次に、ｍビットの各部分データＡ（ｉ，ｊ）を非配列方向、及び配列方向に演算することによって、各行及び各列の配列情報を近似的に表すデータを算出する。このためには、ｍビットのデータを加減乗除の対象にできる体（Field）が必要であり、本発明ではそのために第１の方法としてガロア体（拡大ガロア体）ＧＦ（2^m）を用いる。ガロア体ＧＦ（2^m）を用いた場合には、各行又は各列毎にｍビットの部分データＡ（ｉ，ｊ）、及び必要に応じてｍビットの係数を用いて所定の加減乗除演算（第１又は第２の演算）を行ったときに得られる一つの情報（これを本発明では「パリティ情報」と呼ぶ。）がｍビットであるため、配列情報を少ないデータ量で簡潔に記録できる利点がある。

２を法とする数（０及び１）で表される体をZ_2とすると、ガロア体ＧＦ（2^m）上の演算は、体Z_2上の係数を持つｍ次の既約多項式ＧＦ（Ｘ）を用いて定義することができる。即ち、２つのｍビットの部分データＡ（ｉ，ｊ），Ａ（ｉ’，ｊ’）をそれぞれ２進数表示で（$a_{m-1}\ a_{m-2}\ \cdots\ a_1\ a_0$），（$b_{m-1}$ ｂ

$_{m-2}$ ･･･ b_1 b_0 ）とすると（a_k, b_k は０又は１）、これらをそれぞれ次のように（ｍ－１）次以下の多項式ＡＦ（Ｘ），ＢＦ（Ｘ）に変換する。

ＡＦ（Ｘ）＝$a_{m-1}\cdot X^{m-1}$ ＋$a_{m-2}\cdot X^{m-2}$ ＋･･･ ＋$a_1\cdot X$ ＋a_0 …（１）

ＢＦ（Ｘ）＝$b_{m-1}\cdot X^{m-1}$ ＋$b_{m-2}\cdot X^{m-2}$ ＋･･･ ＋$b_1\cdot X$ ＋b_0 …（２）

この場合、ガロア体ＧＦ（2^m）上でＡＦ（Ｘ）とＢＦ（Ｘ）とを加算する場合には、Ｘの各次数ｋ（ｋ＝０～（ｍ－１））において、係数a_kと係数b_kとを体Z_2上で加算すればよい。体Z_2上では加算と減算とは同じ結果になる。この結果、得られた多項式の係数を２進数表示で表したもの（ベクトル表示）が、部分データＡ（ｉ，ｊ），Ａ（ｉ’，ｊ’）をガロア体ＧＦ（2^m）上で加算した結果になる。これは、ビット毎に排他的論理和演算を行うのと同じ結果である。

次に、ガロア体ＧＦ（2^m）上でＡＦ（Ｘ）にＢＦ（Ｘ）を乗算する場合には、先ず通常の乗算を行って積を求めた後、この積を既約多項式ＧＦ（Ｘ）で除算した後の余りの多項式ＣＦ（Ｘ）を次のように求める（c_k は０又は１）。これを既約多項式ＧＦ（Ｘ）を法（modulus)とする乗算と呼ぶ。この際にもＸの各次数での係数の加算（減算）は体Z_2上で行われる。

ＣＦ（Ｘ）＝$c_{m-1}\cdot X^{m-1}$ ＋$c_{m-2}\cdot X^{m-2}$ ＋･･･ ＋$c_1\cdot X$ ＋c_0 …（３）

この多項式ＣＦ（Ｘ）の係数を２進数表示で表したもの（c_{m-1} c_{m-2} ･･･ c_1 c_0）が、部分データＡ（ｉ，ｊ），Ａ（ｉ’，ｊ’）をガロア体ＧＦ（2^m）上で乗算した結果になる。また、任意のｍビットの係数をβとすると、係数βも（２）式と同様の（ｍ－１）次以下の多項式ＤＦ（Ｘ）で表される。従って、例えば部分データＡ（ｉ，ｊ）に係数βを乗算する場合には、（１）式の多項式ＡＦ（Ｘ）と多項式ＤＦ（Ｘ）との積を既約多項式ＧＦ（Ｘ）を法として計算すればよい。また、例えば部分データＡ（ｉ，ｊ）を係数βで除算する場合には、部分データＡ（ｉ，ｊ）にβの逆元β^{-1}を乗算すればよい。

従って、ｍビットの全てのデータ（全ての部分データＡ（ｉ，ｊ）が含まれる）は、ガロア体ＧＦ（2^m）上のベクトル表示での元とみなすことができ、ｍビットのデータは、多項式表示では、（１）式のような（ｍ－１）次以下の多項式で表すことができる。また、生物学的物質の配列（文字列）を部分データＡ（ｉ，ｊ）に対応させる変換テーブル（所定の規則）の逆変換を用いて、必要に

応じてそのベクトル表示のｍビットのデータを文字列に変換することによって、そのデータに対応する生物学的物質の配列が得られる。

そして、本発明では、例えば図８に示すように、部分データＡ（ｉ，ｊ）が配列方向にＮ個（ｉ＝１～Ｎ）で、非配列方向にＭ個（ｊ＝１～Ｍ）で配列され、各行毎に第１組のパリティ情報（Ｂ１（ｉ），Ｂ２（ｉ），Ｂ３（ｉ））が得られ、各列毎に第２組のパリティ情報（Ｃ１（ｊ），Ｃ２（ｊ），Ｃ３（ｊ））が得られる。これら２組のパリティ情報の内の１つのパリティ情報（例えばＢ１（１））はそれぞれ１つの部分データＡ（ｉ，ｊ）と同じｍビットのデータで表される。

この場合の部分データＡ（ｉ，ｊ）の全体のデータ量ＤＴ１は、以下のようになる。

ＤＴ１＝ｍ・Ｎ・Ｍ（ビット）　　…（４）

また、第１組及び第２組のパリティ情報が、それぞれｅ個（ｅは１以上の整数）のパリティ情報を含むとすると、パリティ情報全体のデータ量ＤＳ１は、以下のようになる。なお、ｅ個のパリティ情報を含む場合には、各行及び各列において、それぞれｅ個までの部分データＡ（ｉ，ｊ）を復元できる。

ＤＳ１＝ｍ・ｅ・（Ｎ＋Ｍ）（ビット）　　…（５）

従って、例えば生物学的物質がＤＮＡを構成するヌクレオチドであるとして、仮にＮ＝６４，Ｍ＝１２８，ｅ＝２とすると、（４）式及び（５）式よりデータ量ＤＴ１，ＤＳ１は以下のようになる。

ＤＴ１＝ｍ・８１９２（ビット）　　…（６）

ＤＳ１＝ｍ・３８４（ビット）≒ＤＴ１／２０　…（７）

従って、パリティ情報のデータ量は、部分データＡ（ｉ，ｊ）全体のデータ量のほぼ１／２０程度に少なくできる。この場合、例えば人間の１本の染色体のＤＮＡの配列は、５０Ｍバイト～２５０Ｍバイト程度のテキストデータで表されるため、予めそのテキストデータを５００個～２５００個程度のブロックに分割し、各ブロック毎に２組のパリティ情報を求めることによって、全部のパリティ情報のデータ量はそのテキストデータのほぼ１／２０程度、即ち２．５Ｍバイト～１２．５Ｍバイト程度に少なくできる。また、その部分データＡ（ｉ，ｊ）が、例

えばテキストデータを１／ｆに小さくしたデータ量である場合には、パリティ情報も更に１／ｆだけ少なくすることができる。

本発明によれば、元の生物学的物質の配列情報を近似的に表す情報（パリティ情報）を、元のテキストデータよりも少ないデータ量のファイルに記録することができる。従って、記録媒体として、ＤＶＤ－ＲＯＭのような大容量の媒体の他に、ＣＤ－ＲＯＭ、及びフラッシュＲＯＭのような小容量でも通常のコンピュータで手軽に再生できる媒体を使用できる。更に、少ないデータ量の配列情報であれば、通信回線を介して短時間に送信できるため、そのパリティ情報は、例えば携帯電話システムなどを介してユーザに安価に供給することも可能となる。

そして、第１組のパリティ情報、及び第２組のパリティ情報を用いることによって、ユーザ側では、２つの生物学的物質の配列の相違する部分を容易に特定することができると共に、相違する部分の個数が各行、又は各列でｅ個以下である場合には、パリティ情報を用いて相違する部分の配列の復元を行うことも可能となる。

なお、テキストデータが記録されたファイルが通常の圧縮技術（ＺＩＰファイル、ＬＨＡファイル等）で圧縮できるように、本発明のパリティ情報が記録されたファイルも更に通常の圧縮技術を用いて圧縮して記録できることは言うまでもない。但し、圧縮されたファイルを使用する場合には、解凍作業が必要になり、最終的には元のファイルを復元する必要があるため、元のファイル自体のデータ量を減らしておくことは極めて有効である。

次に、上記の本発明において、そのガロア体ＧＦ（2^m）上の生成元をαとしたとき、一例として、その第１組のパリティ情報は、複数行の各行のその部分データ（Ａ（ｉ，ｊ））にそれぞれその非配列方向に順次α^{sp}，$\alpha^{s(p+1)}$，$\alpha^{s(p+2)}$，…，$\alpha^{s(p+dp)}$（ｓは０以上の整数、ｐは０以上の整数、ｄｐは１以上の整数）を乗算した後、この演算で得られた複数の積について各行毎に求められた和を含み、その第２組のパリティ情報は、複数列の各列のその部分データ（Ａ（ｉ，ｊ））にそれぞれその配列方向に順次α^{tq}，$\alpha^{t(q+1)}$，$\alpha^{t(q+2)}$，…，$\alpha^{t(q+dq)}$（ｔは０以上の整数、ｑは０以上の整数、ｄｑは１以上の整数）を乗算した後、この演算で得られた複数の積について各列毎に求められた和を含むものである。

この場合、ｐ＝ｑ＝０とすると、第１組のパリティ情報Ｂ１（ｉ）、及び第２組のパリティ情報Ｃ１（ｊ）は、それぞれガロア体ＧＦ（2^m）上の次の演算によって計算される。（８）式のΣはｊについて１～Ｍまでの和を表し、（９）式のΣはｉについて１～Ｎまでの和を表している。

$$B1(i)=\Sigma\alpha^{s(j-1)}\cdot A(i,j)=A(i,1)+\alpha^{s}\cdot A(i,2)+\alpha^{2s}\cdot A(i,3)+\cdots+\alpha^{(M-1)s}\cdot A(i,M) \quad \cdots(8)$$

$$C1(j)=\Sigma\alpha^{t(i-1)}\cdot A(i,j)=A(1,j)+\alpha^{t}\cdot A(2,j)+\alpha^{2t}\cdot A(3,j)+\cdots+\alpha^{(N-1)t}\cdot A(N,j) \quad \cdots(9)$$

そして、（８）式、（９）式でｓ＝ｔ＝０とすると、パリティ情報Ｂ１（ｉ），Ｃ１（ｊ）は、それぞれ部分データＡ（ｉ，ｊ）のガロア体ＧＦ（2^m）上の和、即ち各行又は各列で部分データＡ（ｉ，ｊ）に排他的論理和演算を施して得られる値を示す。従って、簡単な演算で、各行及び各列の配列の近似的な情報を求めることができる。但し、この場合には、各行又は各列で２つの部分データＡ（ｉ，ｊ）が入れ替わったような配列に対しても、パリティ情報Ｂ１（ｉ），Ｃ１（ｊ）は同じ値になってしまう。

次に、ｓ＝ｔ＝１とすると、パリティ情報Ｂ１（ｉ），Ｃ１（ｊ）は、それぞれ各行又は各列で部分データＡ（ｉ，ｊ）に１，α，α^2，α^3，…を乗算して得られる積の和を示す。この場合、各行又は各列で２つの部分データＡ（ｉ，ｊ）が入れ替わったような配列に対しても、パリティ情報Ｂ１（ｉ），Ｃ１（ｊ）は異なった値となるため、例えば２つの生物学的物質の配列間の相違する部分をより正確に特定することができる。そして、或るｓ（≠０）（又はｔ（≠０））の値において、そのように各行又は各列で２つの部分データに常に異なる係数を乗ずるためには、係数$\alpha^{s(j-1)}$（又は$\alpha^{t(i-1)}$）は互いに異なる必要がある。そのためには、αをガロア体ＧＦ（2^m）上の生成元として、各行及び各列の部分データＡ（ｉ，ｊ）の個数を（2^m-1）／ｓ（又は（2^m-1）／ｔ）以下とすればよい。即ち、αを生成元とすることによって、扱う生物学的物質の配列を最も大きくすることができる。

また、各行及び各列において、それぞれ１つのパリティ情報を用いることによって、２つの生物学的物質の配列を比較する場合に、各行及び各列における１つ

の部分データＡ（ｉ，ｊ）の相違部を正確に復元することができる。従って、例えば遺伝子中の一つの塩基（ヌクレオチド）だけが異なるＳＮＰ（一塩基変位多型：Single Nucleotide Polymorphism）は本発明によって容易に検出できると共に、それに対応する正常な配列も容易に復元できる。

更に、各行及び各列における複数個ｓ’及びｔ’の部分データＡ（ｉ，ｊ）の相違部を正確に復元するためには、その第１組のパリティ情報（Ｂ１（ｉ），Ｂ２（ｉ），Ｂ３（ｉ））は、その複数行の各行毎にその整数ｓについて互いに異なる複数（ｓ’個）の値で求めた複数の和を含み、その第２組のパリティ情報（Ｃ１（ｊ），Ｃ２（ｊ），Ｃ３（ｊ））は、その複数列の各列毎にその整数ｔについて互いに異なる複数（ｔ’個）の値で求めた複数の和を含むようにすればよい。その相違部を復元するためには、ガロア体ＧＦ（2^m）上でｓ’元（ｔ’元）の１次連立方程式を解けばよい。

また、本発明においては、その部分データのその配列方向の個数を、その部分データのその非配列方向の個数よりも少なくして、その第２組のパリティ情報の個数を、その第１組のパリティ情報の個数よりも少なくしてもよい。

特に本発明によって得られるパリティ情報をディスプレイに表示するような場合には、その部分データの配列方向の個数をそのディスプレイの横幅に対応する数に制限し、その部分データの非配列方向の個数を大きくすることによって、その非配列方向のデータは、その表示画面上での上下方向へのスクロールによって容易に表示できるため、配列情報を効率的に、且つ分かり易く表示できる。

但し、このように部分データの配列方向の個数が非配列方向の個数よりも少ないときには、その第１組及び第２組のパリティ情報を同じ量にすると、パリティ情報の情報量が全体として多くなる。そこで、配列方向のパリティ情報（第２組のパリティ情報）の個数を非配列方向のパリティ情報（第１組のパリティ情報）の個数よりも少なくすることによって、パリティ情報を少なくして、且つ２つの配列間の相違部の復元を効率的に行うことができる。

また、その部分データの前記非配列方向の個数を、（2^m －１）／４　以下にすることが望ましい。これによって、その非配列方向において４つの異なる係数（α^k，α^{2k}，α^{3k}，α^{4k}）を乗ずることができるため、２つの配列間の各行の

相違部（非配列方向の相違部）を４個まで正確に復元することができる。これは通常のＳＮＰの検出などには十分であると思われる。

また、本発明において、そのガロア体ＧＦ（2^m）を規定する整数mは６４の倍数であることが望ましい。最近のコンピュータにはデータの処理単位が６４ビットであるタイプが増加しているため、整数mを６４の倍数とすることによって、効率的にパリティ情報を算出することができる。

また、本発明において、その生物学的物質の配列を基準配列として、この基準配列のその２組のパリティ情報に対応させて、検査対象の生物学的物質の配列についてその２組のパリティ情報を求め、その４組のパリティ情報よりその基準配列に対するその検査対象の生物学的物質の配列の相違部を求めるようにしてもよい。このように基準配列の２組のパリティ情報と、検査対象の２組のパリティ情報とを比較するのみで、相違部の位置を容易に検出できると共に、相違部が各行及び各列で所定個数以下であれば、４組のパリティ情報と検査対象の生物学的物質の配列とから、連立方程式を解くことによって、基準配列の相違部のデータを正確に復元することもできる。

次に、本発明の生物学的物質の配列情報の記録装置は、その生物学的物質の配列情報を読み取る配列読み取り装置（４）と、その生物学的物質の配列に対応するテキストデータ、又はこのテキストデータを所定の規則に従って変換して得られる数値データを、その生物学的物質の配列方向に複数行で、かつその配列方向に交差する非配列方向に複数列の長さがmビット（mは１６以上の整数）の部分データ（Ａ（ｉ，ｊ））に分割するデータ配列手段（１０，ステップ１０５）と、複数行のその部分データに各行毎にその非配列方向にガロア体ＧＦ（2^m）上の第１の演算を施して第１組のパリティ情報を求めると共に、複数列のその部分データに各列毎にその配列方向にガロア体ＧＦ（2^m）上の第２の演算を施して第２組のパリティ情報を求める演算手段（１０，ステップ１０６）と、その第１組及び第２組のパリティ情報を記録媒体に記録する記録手段（１５）とを有するものである。これによって、本発明のヌクレオチドやアミノ酸などの生物学的物質の配列情報の記録方法が実施できる。

この本発明の記録装置において、そのガロア体ＧＦ（2^m）上の生成元をαと

したとき、一例としてその第１組のパリティ情報は、複数行の各行のその部分データ（Ａ（ｉ，ｊ））にそれぞれその非配列方向に順次α^{sp}，$\alpha^{s(p+1)}$，$\alpha^{s(p+2)}$，…，$\alpha^{s(p+dp)}$（ｓは０以上の整数、ｐは０以上の整数、ｄｐは１以上の整数）を乗算した後、この演算で得られた複数の積について各行毎に求められた和を含み、その第２組のパリティ情報は、複数列の各列のその部分データにそれぞれその配列方向に順次α^{tq}，$\alpha^{t(q+1)}$，$\alpha^{t(q+2)}$，…，$\alpha^{t(q+dq)}$（ｔは０以上の整数、ｑは０以上の整数、ｄｑは１以上の整数）を乗算した後、この演算で得られた複数の積について各列毎に求められた和を含むものである。この場合、部分データ（Ａ（ｉ，ｊ））に乗ずる係数は、生成元αだけから計算できるため、演算が単純化される。

また、本発明の記録媒体は、生物学的物質の配列情報を記録したコンピュータ読み取り可能な記録媒体（１６）であって、その生物学的物質の配列に対応するテキストデータ、又はこのテキストデータを所定の規則に従って変換して得られる数値データを、その生物学的物質の配列方向に複数行で、かつその配列方向に交差する非配列方向に複数列の長さがｍビット（ｍは１６以上の整数）の部分データに分割し、複数行のその部分データに各行毎にその非配列方向にガロア体ＧＦ（2^m）上の第１の演算を施して第１組のパリティ情報を求めると共に、複数列のその部分データに各列毎にその配列方向にガロア体ＧＦ（2^m）上の第２の演算を施して第２組のパリティ情報を求め、その生物学的物質の配列に関する情報を、その第１組及び第２組のパリティ情報として記録したものである。

本発明によれば、ヌクレオチドやアミノ酸などの生物学的物質の配列情報を近似的に表すパリティ情報を少ないデータ量でその記録媒体に記録できるため、記録媒体としてＣＤ－ＲＯＭ，ＣＤ－Ｒ、フラッシュＲＯＭなどの記録容量は比較的少ないが、使い勝手の良い媒体をも使用できる。

この場合、その生物学的物質の配列に対応するそのテキストデータ、又はこのテキストデータに対応するその数値データの４０ビット以上の長さの数学的な要約値（message digest）を更にその記録媒体に記録することが望ましい。

その数学的な要約値は、その生物学的物質の配列に対応するテキストデータ又は数値データに例えばＭＤ５ハッシュ関数（要約値は１２８ビット）、又はＳＨ

Ｓ（Secure Hash Standard）ハッシュ関数（要約値は１６０ビット）などのハッシュ関数の演算を施して得られるものである。その要約値を用いることによって、比較対象の２つの生物学的物質の膨大な配列が一致するかどうかを高い確率で極めて容易に確認することができる。また、パリティ情報を用いて相違する部分データの復元を行った後に、要約値を比較することによって、データが完全に復元できたかを確認することができる。その要約値が４０ビット以上であれば、例えば全人類のＤＮＡのヌクレオチドの配列情報をほぼ互いに衝突することなく表すことができる。

この場合、ガロア体ＧＦ（2^m）を決定する整数ｍが６４の倍数であるときには、要約値が６４ビットの倍数となるハッシュ関数（例えばＭＤ５ハッシュ関数）を用いることが望ましい。演算を効率的に実行できるからである。

次に、本発明の第１の生物学的物質の配列情報の供給方法は、その生物学的物質の配列に対応するテキストデータ、又はこのテキストデータを所定の規則に従って変換して得られる数値データを保持する供給者（２Ａ）が、そのテキストデータ、又はこれに対応するその数値データを第１ファイル（１９）に記録して保持する第１ステップ（ステップ１０４）と、その第１ファイルに記録されているそのテキストデータ、又はこのテキストデータに対応するその数値データを、その生物学的物質の配列方向に複数行で、かつその配列方向に交差する非配列方向に複数列の長さがｍビット（ｍは１６以上の整数）の部分データに分割し、複数行のその部分データに各行毎にその非配列方向にガロア体ＧＦ（2^m）上の第１の演算を施して第１組のパリティ情報（Ｂ１（ｉ）～Ｂ３（ｉ））を求めると共に、複数列のその部分データに各列毎にその配列方向にガロア体ＧＦ（2^m）上の第２の演算を施して第２組のパリティ情報（Ｃ１（ｊ）～Ｃ３（ｊ））を求める第２ステップ（ステップ１０５，１０６）と、その供給者が、その第１組及び第２組のパリティ情報を第２ファイル（２０）に記録して保持する第３ステップ（ステップ１０７）と、その生物学的物質の配列情報のユーザ（２Ｂ）が、通信回線（１）を介してその供給者よりその第２ファイルに記録されているその２組のパリティ情報を受け取る第４ステップ（ステップ１１０，１２９）とを有するものである。

この供給方法は、上記の本発明の生物学的物質の配列情報の記録方法を、その配列情報を供給（販売）する際のビジネスモデルに適用したものである。即ち、本発明のビジネスモデルでは、或る生物ＸのＤＮＡのヌクレオチド、又はタンパク質のアミノ酸などの生物学的物質の配列を最初に解読した供給者は、その配列のテキストデータ（又はこれを変換した数値データ）より、その配列情報を少ないデータ量で近似するパリティ情報を算出し、これをその通信回線を介してユーザに供給する。上述のように、一例としてパリティ情報は、元のテキストデータの１／２０程度のデータ量であるため、そのパリティ情報はその通信回線を介して短時間で受信することができる。

本発明の供給方法においては、更にそのユーザが、その２組のパリティ情報に基づいて検査対象の生物学的物質の配列情報の内のその供給者の生物学的物質の配列情報との相違部を特定する第５ステップ（ステップ１３０，１３１）と、この相違部の配列の復元ができない場合に、そのユーザがその通信回線を介してその供給者よりその第１ファイルに記録されているそのテキストデータ、又はその数値データの内のその配列の復元ができない部分の配列情報を受け取る第６ステップ（ステップ１３５）とを有することが望ましい。

このようにユーザ側で、パリティ情報だけで検査対象の配列内の供給者の配列との相違部の特定、及び復元ができる場合には、それ以上の配列情報を購入する必要が無い。一方、相違部が多く存在し、パリティ情報のみでは全部の正確なデータが復元できない場合には、例えば復元できない部分のテキストデータ（又は数値データ）のみを購入することによって、通信回線を介して必要な配列情報を短時間に購入できる。従って、通信回線として、携帯電話システムのような比較的低速の通信回線も使用できる。

また、本発明の供給方法においては、そのガロア体ＧＦ（2^m）上の生成元をαとしたとき、一例として、その第１組のパリティ情報は、複数行の各行のその部分データにそれぞれその非配列方向に順次α^{sp}，$\alpha^{s(p+1)}$，$\alpha^{s(p+2)}$，…，$\alpha^{s(p+dp)}$（ｓは０以上の整数、ｐは０以上の整数、ｄｐは１以上の整数）を乗算した後、この演算で得られた複数の積について各行毎に求められた和を含み、その第２組のパリティ情報は、複数列の各列のその部分データにそれぞれその配列方

向に順次α^{tq}，$\alpha^{t(q+1)}$，$\alpha^{t(q+2)}$，…，$\alpha^{t(q+dq)}$（ｔは０以上の整数、ｑは０以上の整数、ｄｑは１以上の整数）を乗算した後、この演算で得られた複数の積について各列毎に求められた和を含むものである。

これらのパリティ情報を用いることによって、そのユーザは、ＳＮＰ（一塩基変位多型）などを容易に検出することができる。

また、その供給方法においては、その供給者は、その生物学的物質の配列の長さの情報、及びその配列を表すテキストデータ又はその数値データの数学的な要約値の情報をその通信回線を介して閲覧可能な状態にしておき、そのユーザは、その第４ステップの前にその通信回線を介してその配列の長さの情報及びその数学的な要約値の情報を閲覧する（ステップ１２１）ことが望ましい。

この場合、その供給者は、その生物Ｘの生物学的物質の配列のテキストデータ（又はこれを変換した数値データ）よりハッシュ関数によって算出した要約値（message digest）を例えばインターネット上で閲覧可能にする。これによって、その供給者は、そのテキストデータ自体を公開することなく、最初にその生物Ｘの生物学的物質の配列を解読したことを主張できる。更に、ユーザが同じ配列情報を異なる供給者から誤って購入することも防止できる。

また、或るユーザが、その供給者よりその生物学的物質の配列情報を購入した後、購入した配列情報よりそのハッシュ関数によって要約値を算出し、その配列の長さも求める。そして、この配列の長さ、及び要約値をインターネット上で公開されている値と比較することによって、購入した配列情報が正確なものであるかどうかを極めて高い確率で確認できる。

この場合、一例として、その数学的な要約値は、４０ビット以上で１９２ビット以下のデータであり、その供給者は、更にその生物学的物質の所定の一部の配列の情報をその通信回線を介して閲覧可能な状態にしておくことが望ましい。その要約値、及びその配列の長さの他に、そのように例えばその配列の先頭の８個程度、及び後端の８個程度の配列を比較することによって、同一性の確認をより高精度に行うことができる。

次に、本発明による第２の生物学的物質の配列情報の記録方法は、その生物学的物質の配列に対応するテキストデータ、又はこのテキストデータを所定の規則

に従って変換して得られる数値データを、その生物学的物質の配列方向に複数行で、かつその配列方向に交差する非配列方向に複数列の長さがｍビット（ｍは１６以上の整数）の部分データ（Ａ（ｉ，ｊ））に分割し、その部分データの最大値をＮｍａｘ、この最大値Ｎｍａｘよりも大きい素数をＰとして、複数行のその部分データに各行毎にその非配列方向にガロア体ＧＦ（Ｐ）上の第１の演算を施して第１組のパリティ情報を求めると共に、複数列のその部分データに各列毎にその配列方向にガロア体ＧＦ（Ｐ）上の第２の演算を施して第２組のパリティ情報を求め、その第１組及び第２組のパリティ情報でその生物学的物質の配列を表すものである。

斯かる本発明によれば、その生物学的物質としては、例えば一列のヌクレオチド又は一列のアミノ酸が考えられる。前者の一列のヌクレオチドは、例えば或るＤＮＡの一方の鎖の少なくとも一部、或るＲＮＡを構成する１列の重合体の鎖の少なくとも一部、又は或る遺伝子の少なくとも一部である。そして、例えば図７の例において、その生物学的物質の配列情報を示すテキストデータを配列方向にＮ個（ｉ＝１～Ｎ）で、非配列方向にＭ個（ｊ＝１～Ｍ）の部分テキストデータＴ（ｉ，ｊ）に分割し、図８に示すように、各部分テキストデータＴ（ｉ，ｊ）をそれぞれｍビットの部分データＡ（ｉ，ｊ）に変換する。ｍビットの部分データＡ（ｉ，ｊ）は、それぞれｎ個（図８の例ではｎ＝１６）の連続する生物学的物質の配列を表している。

次に、ｍビットの各部分データＡ（ｉ，ｊ）を非配列方向、及び配列方向に演算することによって、各行及び各列の配列情報を近似的に表すデータを算出する。このためには、ｍビットのデータを加減乗除の対象にできる体（Field）が必要であり、本発明ではそのために第２の方法としてガロア体ＧＦ（Ｐ）を用いる。ガロア体ＧＦ（Ｐ）を用いた場合、その部分データＡ（ｉ，ｊ）の最大値が（2^m－１）であると、その素数Ｐは（ｍ＋１）ビットの値となり、各行及び各列の部分データにガロア体ＧＦ（Ｐ）上の所定の演算を施して得られる一つの情報（これを本発明では「パリティ情報」と呼ぶ。）も（ｍ＋１）ビットで表されるため、各パリティ情報のデータ量が１ビットだけ大きくなる。しかしながら、全体として見ると、ガロア体ＧＦ（2^m）を用いる場合とほぼ同程度に少ないパリティ情

報で、元の配列情報を近似することができる。更に本発明によれば、そのパリティ情報を求めるための演算はガロア体ＧＦ（2^m）での演算に比べると簡単である。

本発明においても、２つの配列の比較を行う際にそのパリティ情報を比較することによって、相違部の特定、及び相違部の或る程度の復元を行うことができる。

本発明においては、その部分データの最大値Ｎｍａｘは（2^m －１）よりも小さい値であることが望ましい。これを実現する最も簡単な方法は、部分データＡ（ｉ，ｊ）として、この部分データに対応する配列を表すテキストデータそのもの（数値データとみなす）を使用することである。このとき、その素数Ｐは、次の関係を満たすように選択できることが望ましい。

$2^m > P > Nmax$　　…（Ａ１）

これは、その素数Ｐをその最大値Ｎｍａｘよりも大きいｍビットの数とすることを意味する。これによって、パリティ情報もｍビットのデータとなるため、ガロア体ＧＦ（2^m）よりも簡単な演算でよいことと、ガロア体ＧＦ（2^m）を用いる場合と同じデータ量でパリティ情報を表すこととを両立できる。

この場合、そのガロア体ＧＦ（Ｐ）上の生成元をδとしたとき、一例としてその第１組のパリティ情報は、複数行の各行のその部分データにそれぞれその非配列方向に順次δ^{sp}，$\delta^{s(p+1)}$，$\delta^{s(p+2)}$，…，$\delta^{s(p+dp)}$（ｓは０以上の整数、ｐは０以上の整数、ｄｐは１以上の整数）を乗算した後、この演算で得られた複数の積について各行毎に求められた和を含み、その第２組のパリティ情報は、複数列の各列のその部分データにそれぞれその配列方向に順次δ^{tq}，$\delta^{t(q+1)}$，$\delta^{t(q+2)}$，…，$\delta^{t(q+dq)}$（ｔは０以上の整数、ｑは０以上の整数、ｄｑは１以上の整数）を乗算した後、この演算で得られた複数の積について各列毎に求められた和を含むものである。その整数ｓ，ｔの値を調整することによって、容易に各行及び各列で複数のパリティ情報を求めることができる。

次に、本発明の第２の生物学的物質の配列情報の供給方法は、その生物学的物質の配列に対応するテキストデータ、又はこのテキストデータを所定の規則に従って変換して得られる数値データを保持する供給者（２Ａ）が、そのテキストデータ、又はこれに対応するその数値データを第１ファイル（１９）に記録して保

持する第１ステップ（ステップ１０４）と、その第１ファイルに記録されているそのテキストデータ、又はこのテキストデータに対応するその数値データを、その生物学的物質の配列方向に複数行で、かつその配列方向に交差する非配列方向に複数列の長さがｍビット（ｍは１６以上の整数）の部分データに分割し、その部分データの最大値をＮｍａｘ、この最大値Ｎｍａｘよりも大きい素数をＰとして、複数行のその部分データに各行毎にその非配列方向にガロア体ＧＦ（Ｐ）上の第１の演算を施して第１組のパリティ情報を求めると共に、複数列のその部分データに各列毎にその配列方向にガロア体ＧＦ（Ｐ）上の第２の演算を施して第２組のパリティ情報を求める第２ステップと、その供給者が、その第１組及び第２組のパリティ情報を第２ファイル（２０）に記録して保持する第３ステップと、その生物学的物質の配列情報のユーザ（２Ｂ）が、通信回線（１）を介してその供給者よりその第２ファイルに記録されているその２組のパリティ情報を受け取る第４ステップとを有するものである。

この供給方法は、上記の本発明の第２の生物学的物質の配列情報の記録方法を、その配列情報を供給（販売）する際のビジネスモデルに適用したものである。即ち、本発明のビジネスモデルでは、或る生物ＸのＤＮＡのヌクレオチド、又はタンパク質のアミノ酸などの生物学的物質の配列を最初に解読した供給者は、その配列のテキストデータ（又はこれを変換した数値データ）より、その配列情報を少ないデータ量で近似するパリティ情報を算出し、これをその通信回線を介してユーザに供給する。本発明においても、ガロア体ＧＦ（2^m）を用いる場合とほぼ同程度にそのパリティ情報は、元のテキストデータに比べて少なくできるため、そのパリティ情報はその通信回線を介して短時間で受信することができる。

本発明の供給方法においても、更にそのユーザが、その２組のパリティ情報に基づいて検査対象の生物学的物質の配列情報の内のその供給者の生物学的物質の配列情報との相違部を特定する第５ステップと、この相違部の配列の復元ができない場合に、そのユーザがその通信回線を介してその供給者よりその第１ファイルに記録されているそのテキストデータ、又はその数値データの内のその配列の復元ができない部分の配列情報を受け取る第６ステップとを有することが望ましい。

このようにユーザ側で、パリティ情報だけで検査対象の配列内の供給者の配列との相違部の特定、及び復元ができる場合には、それ以上の配列情報を購入する必要が無い。一方、相違部が多く存在し、パリティ情報のみでは全部の正確なデータが復元できない場合には、例えば復元できない部分のテキストデータ（又は数値データ）のみを購入することによって、通信回線を介して必要な配列情報を短時間に購入できる。従って、通信回線として、携帯電話システムのような比較的低速の通信回線も使用できる。

図面の簡単な説明

図１は、本発明の実施の形態の一例で使用されるコンピュータシステムを示す概略構成図である。図２は、その実施の形態の一例で処理対象とするＤＮＡ、及びそのヌクレオチドの配列のバイナリーデータによる表現の例を示す図である。図３は、その実施の形態の一例におけるＤＮＡ情報の供給者の動作の一部を示すフローチャートである。図４は、図３の動作に続くＤＮＡ情報の供給者の動作を示すフローチャートである。図５は、その実施の形態の一例におけるＤＮＡ情報のユーザの動作の一部を示すフローチャートである。図６は、図５の動作に続くＤＮＡ情報のユーザの動作を示すフローチャートである。図７は、標準試料Ｅ（ＤＮＡ）のヌクレオチド（２０４８個）の配列を表すテキストデータを４行で３２列の部分テキストデータＴ（ｉ，ｊ）に分割した状態を示す図である。図８は、標準試料Ｅの部分データＡ（ｉ，ｊ）、及びこれらから算出されるパリティＢ１（ｉ）～Ｃ３（ｊ）を示す図である。図９は、試料Ｆ（ＤＮＡ）のヌクレオチド（２０４８個）の配列を表すテキストデータを４行で３２列の部分テキストデータＴＦ（ｉ，ｊ）に分割した状態を示す図である。図１０は、試料Ｆの部分データＡＦ（ｉ，ｊ）、及びこれらから算出されるパリティＢ１Ｆ（ｉ）～Ｃ３Ｆ（ｊ）を示す図である。図１１は、標準試料Ｅのパリティと異なる試料Ｆのパリティ、及び復元された部分データを示す図である。図１２は、未知数Ｘ１，Ｘ２，Ｙ１，Ｙ２をガロア体ＧＦ（2^{128}）上で求める場合の計算を示すフローチャートである。図１３は、図７の標準試料Ｅのヌクレオチドの配列を表すテキストデータをバイナリーデータに変換した後、５行で１３列の部分データＢ（ｉ，

ｊ）に分割した状態を示す図である。図１４は、試料Ｇ（タンパク質）のアミノ酸（８２０個）の配列を表すテキストデータを４行で２６列の部分テキストデータに分割した状態を示す図である。図１５は、図１４のアミノ酸の配列に対して計算されたパリティＢ１Ｇ（ｉ）～Ｃ３Ｇ（ｊ）を示す図である。

発明を実施するための最良の形態

以下、本発明の好ましい実施の形態の一例につき図面を参照して説明する。本例は、所定のＤＮＡ（デオキシリボ核酸：deoxyribonucleic acid)のヌクレオチド（生物学的物質）の配列情報をコンピュータシステムで処理する場合に、本発明を適用したものである。

図１は、本例のコンピュータシステム２Ａの概略構成を示し、この図１において、コンピュータシステム２Ａの中心は、ＣＰＵ（中央演算処理ユニット）、ＲＡＭ，ＲＯＭ等のメモリ、及びハードディスク装置等の記憶装置等からなる情報処理装置１０である。情報処理装置１０には、ビデオＲＡＭ（ＶＲＡＭ）１１を介してＣＲＴディスプレイよりなる表示装置１２が接続されると共に、Ｉ／Ｏユニット（入出力装置）１４を介して、記録可能なＣＤ－Recordableディスク（以下、「ＣＤ－Ｒ」と言う）１６に対するデータの書き込み、及びＣＤ－ＲやＣＤ－ＲＯＭからのデータの読み込みを行うことができるＣＤ－Ｒ／ＲＷドライブ１５が接続されている。情報処理装置１０には、Ｉ／Ｏユニット１４を介して更に大容量の記憶装置としての記憶容量が数１００Ｇバイト程度の磁気ディスク装置１７が接続されている。

本例の情報処理装置１０中のハードディスク装置には、予めＣＤ－Ｒ／ＲＷドライド１５を介してオペレーティングシステム、及び後述のようにＤＮＡの配列情報を処理するためのアプリケーション・プログラムがインストールされている。また、ＣＤ－Ｒ１６が本発明の記録媒体に対応しているが、記録媒体としては、ＣＤ－ＲやＣＤ－ＲＯＭの他に、フラッシュＲＯＭ、フレキシブルディスク、光磁気ディスク（ＭＯ）、デジタルビデオディスク（ＤＶＤ）、又はハードディスク装置（例えばインターネットを介して接続できるサーバに備えられたもの）等を使用することができる。

情報処理装置１０には更に、文字情報の入力装置としてのキーボード１３、ポインティング・デバイス（入力装置）としての光学式のマウス２０４、及びルータ（又はモデム等でもよい）よりなる通信制御ユニット１８が接続されている。マウス２０４は、表示装置１２の表示画面上のカーソルの位置を指定する信号を発生する変位信号発生部２０７、選択すべき情報を指定する信号や各種コマンド等を発生するための左スイッチ２０４ａ及び右スイッチ２０４ｂ（信号発生装置）を備えている。情報処理装置１０、ＶＲＡＭ１１、表示装置１２、キーボード１３、マウス２０４、Ｉ／Ｏユニット１４、ＣＤ－Ｒ／ＲＷドライブ１５、磁気ディスク装置１７、及び通信制御ユニット１８等よりコンピュータシステム２Ａが構成されている。オペレーティングシステムとして本例ではＷｉｎｄｏｗｓ（Microsoft Corporation の登録商標）を使用している。なお、オペレーティングシステムとして、それ以外のＵＮＩＸ（X/Open の登録商標）、ＯＳ／２（IBM Corporatin の登録商標）、ＭａｃＯＳ（Apple Computer の登録商標）、又はＬｉｎｕｘ（Linus Torvalds の商標又は登録商標）等を使用する場合にも本発明が適用できることは言うまでもない。

そして、コンピュータシステム２Ａ（情報処理装置１０）は、通信制御ユニット１８を介して一般電話回線よりなる通信ネットワーク１に接続され、通信ネットワーク１には各種コンテンツのプロバイダ３、及び別のコンピュータシステム２Ｂ、及び不図示の多くのサーバやコンピュータシステムが接続されている。また、本例のコンピュータシステム２Ａ，２Ｂ及びプロバイダ３は、通信ネットワーク１を介するインターネットによって相互に接続されている。この場合、コンピュータシステム２Ａの所有者がＤＮＡ情報の供給者（販売者）であり、コンピュータシステム２Ｂの所有者がそのＤＮＡ情報のユーザ（購入者）である。そして、後者のコンピュータシステム２Ｂには、予め前者のコンピュータシステム２Ａと同様のＤＮＡの配列情報を処理するためのアプリケーション・プログラムがインストールされている。

さて、本例のコンピュータシステム２Ａの情報処理装置１０には、Ｉ／Ｏユニット１４を介して、生物学的物質としてのＤＮＡ中の一列のヌクレオチドの配列（又は塩基の配列）を読み取るための配列読み取り装置としてのシーケンサー

(DNA Sequencer) ４が接続されている。シーケンサー４は、一例としてサンガーの方法 (Sanger method) によってＤＮＡを構成する１対の重合体の鎖の一方の鎖のヌクレオチドの配列を読み取る。サンガーの方法は、例えば文献１ (Maxim D. Frank-Kamenetskii: Unraveling DNA (the most important molecule of life, revised and updated), translated by Lev Liapin, Chapter 6 (pp. 59-70) (Perseus Books, 1997)) に開示されている。シーケンサー４は、読み取った一列のヌクレオチドの配列をテキストデータ形式で内部の大容量の記憶装置に記憶すると共に、情報処理装置１０からの要求に応じて、その記憶装置中の所定のヌクレオチドの配列のテキストデータをＩ／Ｏユニット１４を介して情報処理装置１０に供給する。これに対して情報処理装置１０は、ＤＮＡの配列情報を処理するためのアプリケーション・プログラムに基づいて以下の処理を行う。なお、シーケンサー４の代わりに、ＤＮＡ及びＲＮＡ（リボ核酸：ribonucleic acid）等の核酸を構成する一列のヌクレオチドの配列（又は塩基の配列）の情報のデータベースを接続してもよい。

先ず、本例の情報処理装置１０の第１の基本的な処理動作につき説明する。情報処理装置１０は、シーケンサー４から供給される所定のＤＮＡのヌクレオチドの配列を示すテキストデータ（本例ではアスキーコード（ＡＮＳＩ形式）を用いる）を磁気ディスク装置１７中のマスターファイル１９にそのまま記録すると共に、そのテキストデータをよりデータ量の少ない数値データに変換し、この変換後の数値データを磁気ディスク装置１７中のワーキングファイル２０に記録する。なお、以下の説明において、２進数表示の数ｋは bin(k) で、１６進数表示の数ｋは hex(k) で表すものとする。

この場合、ＤＮＡは４種類のヌクレオチドより構成されており、シーケンサー４から供給されるテキストデータ中では、塩基としてアデニン（adenine）を含むヌクレオチド、グアニン（guanine）を含むヌクレオチド、シトシン（cytosine）を含むヌクレオチド、及びチミン（thymine）を含むヌクレオチドがそれぞれ文字Ａ，Ｇ，Ｃ，及びＴで表されている。そして、文字Ａ，Ｇ，Ｃ，及びＴには、データ上ではそれぞれhex(41), hex(47), hex(43), hex(54)よりなる１バイト（８ビット）のアスキーコードが割り当てられている。また、ＲＮＡの場合には、チミン

を含むヌクレオチドの代わりにウラシル（uracil）を含むヌクレオチドが、文字U（hex(55)）で表されている。従って、ｎ個のヌクレオチドの配列を示すテキストデータのデータ量はｎバイトとなる。なお、それらのｎ個のヌクレオチドの配列は、ｎ個の塩基（アデニン、グアニン、シトシン、チミン（又はウラシル））の配列ともみなすことができる。

本例ではそのテキストデータを、情報量を少なくすることなく最も少ないデータ量で表すために、ＤＮＡ中の４種類のヌクレオチドを互いに異なる２ビットのデータで表す。この際に、ＤＮＡにおいては、１対の塩基（アデニン及びチミン）が互いに相補的であり、別の１対の塩基（グアニン及びシトシン）が互いに相補的である。そこで、相補的な塩基を含む１対のヌクレオチドを互いに相補的であるとして、１対の互いに相補的なヌクレオチド、即ちアデニンを含むヌクレオチド及びチミンを含むヌクレオチドに、互いにビット反転の関係にある１対のデータを割り当て、別の１対の互いに相補的なヌクレオチド、即ちグアニンを含むヌクレオチド及びシトシンを含むヌクレオチドに、互いにビット反転の関係にある別の１対のデータを割り当てる。本例ではそのデータの割り当てとして表１（変換テーブル）を用いる。なお、表１は、ヌクレオチドの配列を示すテキストデータ中の文字Ａ，Ｔ（又はU），Ｇ，Ｃ，をそれぞれbin(00), bin(11), bin(01), bin(10) で置換することを意味している。

《表１》

ヌクレオチド	２ビットのデータ
アデニンを含むヌクレオチド（Ａ）	bin(00)
チミン（ウラシル）を含むヌクレオチド（Ｔ又はU）	bin(11)
グアニンを含むヌクレオチド（Ｇ）	bin(01)
シトシンを含むヌクレオチド（Ｃ）	bin(10) 。

なお、本例では各ヌクレオチドを２ビットのデータで表しているが、これは各塩基を２ビットのデータで表すのと等価である。また、データの割り当ては表１には限定されず、例えばチミンを含むヌクレオチドをbin(00) 、アデニンを含むヌクレオチドをbin(11) とするか、又はグアニンを含むヌクレオチドをbin(10) 、シトシンを含むヌクレオチドをbin(01) としてもよい。それ以外に、アデニンを

含むヌクレオチド及びチミンを含むヌクレオチドに、１対のデータbin(01), bin(10) を割り当て、グアニンを含むヌクレオチド及びシトシンを含むヌクレオチドに１対のデータbin(00), bin(11) を割り当てるようにしてもよい。また、ＲＮＡの場合には、チミンを含むヌクレオチドに割り当てられているデータをウラシルを含むヌクレオチドに割り当てて、それ以外のヌクレオチドにはＤＮＡのヌクレオチドと同じデータを割り当てればよい。

本例では図２に示すＤＮＡ分子５のヌクレオチドの配列情報を扱うものとする。その配列情報は、ＮＣＢＩ（The National Center for Biotechnology Information ）より提供されているウェブサイト１（ftp://ncbi.nlm.nih.gov/genbank/genomes/bacteria/）より入手した大腸菌（Escherichia coli: E. coli）のＤＮＡの一列のヌクレオチドの配列の一部である。

図２において、ＤＮＡ分子５は、１対の重合体の鎖６Ａ，６Ｂ（二重らせん）より構成され、一方の重合体の鎖６Ａは、アデニンを含むヌクレオチド７Ａ、グアニンを含むヌクレオチド７Ｇ、シトシンを含むヌクレオチド７Ｃ、及びチミンを含むヌクレオチド７Ｔよりなる４種類のヌクレオチドの配列であり、他方の重合体の鎖６Ｂは、鎖６Ａに対して相補的なヌクレオチドの配列である。この際に、図１の情報処理装置１０には一方の重合体の鎖６Ａの配列を示すテキストデータ、即ち”ＡＧＣＴＴＴ・・・”の文字列のデータが供給される。それに対して、情報処理装置１０は、そのテキストデータ中の文字Ａ，Ｇ，Ｃ，Ｔを表１の変換テーブルに基づいて順次２ビットのデータに変換することによって、数値データとしてのバイナリーデータＢＮＡ（＝bin(0001101111･･･)）を得る。そして、このバイナリーデータＢＮＡが図１の磁気ディスク装置１７のワーキングファイル２０に記録される。そのバイナリーデータＢＮＡのデータ量は、元のテキストデータの１／４となっている。

この場合、そのワーキングファイル２０の先頭の所定数のバイトの領域に、例えばその配列がＤＮＡ又はＲＮＡのどちらかを示すデータ（即ち、bin(11) を文字Ｔ又は文字Ｕの何れに解釈するかを示すデータ）、ヌクレオチドの個数を示すデータ、及びその他の必要なデータを記録しておけばよい。また、そのワーキングファイル２０の長さが１バイト（８ビット）単位で規定されている場合に、バ

イナリーデータＢＮＡの末尾で１バイトの端数のデータが生じたときには、予め定めておいたダミーデータを付加すればよい。これでもデータ量は殆ど増加しない。そして、一例としてユーザ（コンピュータシステム２Ｂの所有者）から供給者（コンピュータシステム２Ａの所有者）に対して図２のＤＮＡ分子５の配列情報の購入希望が届いたときに、ワーキングファイル２０のデータが通信ネットワーク１及び不図示のプロバイダを介して、電子メールの添付ファイルとしてコンピュータシステム２Ｂ側に供給される。この際に、そのワーキングファイル２０のデータを更に圧縮ファイル（ＺＩＰファイル、又はＬＨＡファイル等）として送信してもよい。この際に、ワーキングファイル２０のデータ量はもとのテキストデータのほぼ１／４であるため、元のテキストデータ（更に圧縮ファイルとした場合も同様）自体を送信する場合に比べて送信時間はほぼ１／４となり、供給者側及びユーザ側双方の通信コストが低減できる。

そして、ユーザ側で、その受信したワーキングファイル２０のデータから図２の一方の重合体の鎖６Ａの配列のテキストデータを復元する場合には、コンピュータシステム２Ｂにおいて、ワーキングファイル２０中のバイナリーデータＢＮＡを、表１を用いて文字Ａ，Ｇ，Ｃ，Ｔ（又はU）の何れかに順次逆変換すればよい。また、その際に例えば図２の他方の相補的な重合体の鎖６Ｂのヌクレオチドの配列を示すテキストデータが必要になった場合には、コンピュータシステム２Ｂにおいて、図２に示すように、バイナリーデータＢＮＡのビット毎の反転操作を行って反転バイナリーデータNOT(BNA)（＝bin(1110010000･･･)）を得る。この反転バイナリーデータNOT(BNA)は、他方の重合体の鎖６Ｂのヌクレオチドの配列を示すテキストデータ（文字列”ＴＣＧＡＡＡ・・・”）を表１に従って変換したバイナリーデータＢＮＢと同一である。従って、その反転バイナリーデータNOT(BNA)を、表１を用いて文字Ａ，Ｇ，Ｃ，Ｔ（又はU）の何れかに順次逆変換するのみで、極めて高速に相補的な重合体の鎖６Ｂの配列のテキストデータを得ることができる。この際に、通常のコンピュータにおいては、ビット毎の反転操作は、極めて高速に実行することができる。なお、そのビット毎の反転操作は、例えばbin(111111･･･)との排他的論理和演算で代用してもよい。

なお、ワーキングファイル２０のデータを通信ネットワーク１を介してユーザ

側に送信する代わりに、ワーキングファイル２０の内容をＣＤ－Ｒ／ＲＷドライブ１５によってＣＤ－Ｒ１６に記録し、このＣＤ－Ｒ１６を郵送等によってユーザ側に供給してもよい。例えば一人の人間の全部のＤＮＡの配列情報（ゲノム）は、テキストデータでは３Ｇバイト程度になるが、これを表１を用いて本例の数値データとしてのバイナリーデータに変換すると、３／４Ｇバイト程度、即ち７５０Ｍバイト程度になる。現在のＣＤ－Ｒ，ＣＤ－ＲＯＭの容量は約６５０Ｍバイトであるため、その７５０Ｍバイト程度のバイナリーデータは例えば一部又は全部を圧縮ファイルとすることによって、余裕を持ってＣＤ－Ｒ１６に記録することができる。これに対して、その７５０Ｍバイト程度のデータを通信ネットワーク１を介して送信しようとすると、現状でも送信時間がかかり過ぎる場合がある。

また、一つのアミノ酸の種類は一列の３個のヌクレオチドの配列、即ち一つの遺伝子コドン（codon ）によって決定される。そこで、１つのアミノ酸に対応する３個のヌクレオチドをそれぞれ２ビットのデータで表したときに得られる６ビットのデータの内で、最も小さいデータでそのアミノ酸を表すようにしてもよい。この際に、個々のデータは、１バイト単位が扱い易いため、その６ビットのデータの前後に２ビットの識別データを付加して得られる１バイトのデータで１つのアミノ酸を表すようにしてもよい。これによって、ヌクレオチドとアミノ酸とで共通のコードを使用できる利点がある。

次に、本例の情報処理装置１０の第２の基本的な処理動作につき説明する。本例では、ヌクレオチドの配列を示す膨大な量のテキストデータ（又はこれを表１に基づいて変換して得られる数値データ）より、所定のハッシュ関数を用いて数学的な要約値（message digest）を算出する。本例ではそのハッシュ関数として、ライベスト（R. Rivest)によって提案されたＭＤ５ハッシュ関数を使用する。ＭＤ５ハッシュ関数のアルゴリズムについては、ネットワークワーキンググループ及びライベストによって開設されているウェブサイト２(http://www.kleinscmidt.com/edi/md5.htm)に開示されている。また、そのＭＤ５ハッシュ関数のアルゴリズムは、国際公開公報ＷＯ01/80431 A1 にも開示されている。また、２００２年１０月１６日付け提出の米国出願第１０／２７２，１０７号で開示されている

ＭＤ５ハッシュ関数及びその他のハッシュ関数のアルゴリズムは、そっくりそのまま引用して本願に組み込まれている。或るテキストデータ（テキストファイル）にそのＭＤ５ハッシュ関数を施すことによって、１２８ビットの要約値が得られる。通常のコンピュータでも今後は処理単位が６４ビットのＣＰＵが使用されるようになると考えられるが、この場合に１２８（＝２・６４）ビットの要約値は非常に扱い易い長さである。この場合には、１９２（＝３・６４）ビットの要約値も比較的扱い易いと考えられる。

また、本例では、そのＭＤ５ハッシュ関数のプログラムとして、そのウェブサイト２において公開されている、ＲＳＡデータセキュリティー社（RSA Data Security Inc.）によって開発されたプログラムを使用した。

その要約値の使用方法の一例として、ＤＮＡの配列情報の供給者（情報処理装置１０）は、所定の生物のＤＮＡのヌクレオチドの配列を読み取り、これに対応するテキストデータより、上記のハッシュ関数を用いて要約値を算出し、この要約値をその生物の名称、及びＤＮＡの位置を示す情報と共にインターネット上で閲覧可能にする。これによって、その供給者は、そのテキストデータ自体を公開することなく、その生物のＤＮＡの配列情報を最先に解読したことを主張できると考えられる。その後、或るユーザからのその配列情報の購入希望が来たときに、その供給者は、そのヌクレオチドの配列のテキストデータを表１を用いてバイナリーデータに変換し、このバイナリーデータを例えば通信ネットワーク１を介して電子メールの形でそのユーザに送信する。これに対してユーザ側では、そのバイナリーデータを表１を用いてテキストデータに逆変換し、この逆変換されたテキストデータに上記のハッシュ関数を施して要約値を求める。

そして、この要約値とその供給者によって公開されている要約値とが等しいときには、購入した配列情報が、供給の保持している配列情報と等しいことが極めて高い確率で保証される。更に、ユーザ側では、複数の供給者が公開している要約値を比較することによって、同じ配列情報を異なる複数の供給者から重複して購入することを防止することができる。これらの際に、ヌクレオチドの配列の長さ、及び先端部や末尾の一部の短い配列の比較を行うことによって、その配列情報の同一性を高めることができる。

なお、ハッシュ関数としては、例えば文献２(FIPS Publication 180, 1993)で開示されているように、ＮＢＳ(National Bureau of Standards)によって提案されたＳＨＳ（Secure Hash Standard）ハッシュ関数を使用してもよい。ＳＨＳハッシュ関数は、ＭＤ５ハッシュ関数よりも複雑な演算を行うと共に、１６０ビットの要約値が得られる。これに関して、例えばタンパク質を構成するアミノ酸の配列数は２０個～１０００個程度であり、特に一文字表記を使用する際にはそれに対応するテキストデータも２０バイト～１ｋバイト程度に短くなるため、要約値から元のテキストデータが推定し易いと考えられる。そこで、アミノ酸の配列情報の要約値を求める際には、ＳＨＳハッシュ関数を使用する方が望ましいことがある。

また、例えばヌクレオチドの配列を示す２つの膨大な長さのテキストデータの同一性を確認するために、ハッシュ関数の要約値を算出するような場合には、それ程複雑な計算を繰り返して行う必要は無いと考えられる。そこで、このような用途では、例えば文献３（R. L. Rivest:"The MD4 message digest algorithm", Lecture Notes in Computer Science, 537,303-311(1991)）で開示されているＭＤ４ハッシュ関数を使用してもよいと考えられる。また、そのように単に同一性を確認する用途では、要約値の長さも４０ビット～１２８ビット程度でよい場合がある。

次に、本例のＤＮＡ情報の供給者（コンピュータシステム２Ａ）と、ユーザ（コンピュータシステム２Ｂ）との間でＤＮＡの配列情報を受け渡す際のビジネスモデルの一例につき図３～図６のフローチャートを参照して詳細に説明する。先ず、ＤＮＡ情報の供給者側では、図３のステップ１０１において、シーケンサー４を使用して標準となる試料（標準試料Ｅとする）のＤＮＡ中の一方の一列のヌクレオチドの配列を読み取り、読み取った配列を表すテキストデータＴＸ１を情報処理装置１０に供給する。本例では、その標準試料Ｅを大腸菌として、そのテキストデータＴＸ１として、図７に示すように、上記のウェブサイト１から入手した大腸菌のＤＮＡの配列情報の内の、最初から２０４８個までのヌクレオチドの配列を示すテキストデータを使用する。

標準試料ＥのＤＮＡ配列は配列番号１に示されている。図７のテキストデータ

は、配列番号１の配列から数字データを除いて、ａ，ｇ，ｃ，ｔの文字をそれぞれＡ，Ｇ，Ｃ，Ｔで置き換えたものに相当する。

次のステップ１０２において、情報処理装置１０は、供給されたテキストデータＴＸ１に上記のＭＤ５ハッシュ関数を施して１２８ビットの要約値ＡＢ１を求めると共に、そのヌクレオチドの配列の数ＮＡ１、及び先頭と末尾との８個ずつのヌクレオチドの配列ＳＴ１，ＳＢ１を求める。テキストデータＴＸ１に対する具体的な値は下記の通りである。

ＡＢ１＝hex(849339ac244cde42b5346ab5989aab61) …（１１）

ＮＡ１＝２０４８

ＳＴ１＝ＡＧＣＴＴＴＴＣ，ＳＢ１＝ＣＧＣＧＡＡＧＧ

次のステップ１０３において、情報処理装置１０は、テキストデータＴＸ１を逆方向に並べ替えたテキストデータＴＸＲ１（＝ＧＧＡＡＧＣ・・・ＴＴＴＣＧＡ）を求め、このテキストデータＴＸＲ１のＭＤ５ハッシュ関数による要約値ＡＢＲ１、及びこのテキストデータＴＸＲ１の先頭と末尾との８個ずつのヌクレオチドの配列ＳＴＲ１，ＳＢＲ１を求める。配列ＳＴＲ１，ＳＢＲ１は、上記の配列ＳＢ１，ＳＴ１をそれぞれ逆方向に並べ替えることによって容易に求めることができる。これらの具体的な値は以下の通りである。

ＡＢＲ１＝hex(4eb1feae30f522642b912ce3ea09652b) …（１２）

ＳＴＲ１＝ＧＧＡＡＧＣＧＣ，ＳＢＲ１＝ＣＴＴＴＴＣＧＡ

次のステップ１０４において、情報処理装置１０は、標準試料Ｅの名前の情報（試料を特定する情報）、配列の数ＮＡ１、テキストデータＴＸ１、配列ＳＴ１，ＳＢ１、要約値ＡＢ１、逆方向の配列ＳＴＲ１，ＳＢＲ１、及び逆方向の要約値ＡＢＲ１を磁気ディスク装置１７のマスターファイル１９に記録する。この際に、マスターファイル１９を複数のファイルとして、テキストデータＴＸ１と、それ以外のデータとを別のファイルに記録してもよい。また、テキストデータＴＸ１が例えば１００Ｍバイト程度以上になる場合には、テキストデータＴＸ１を複数のマスターファイルに分割して記録してもよい。

次のステップ１０５において、情報処理装置１０は、図７に示すように、標準試料Ｅのテキストデータ ＴＸ１を配列方向（ヌクレオチドの配列方向）にＮ行で、

その配列方向に直交する方向（以下、「非配列方向」という）にM列の１６文字の長さの部分テキストデータＴ（ｉ，ｊ）（ｉ＝１～Ｎ，ｊ＝１～Ｍ）に分割する。なお、Ｎ，Ｍはそれぞれ２以上の任意の整数であり、（４）式、（５）式を用いて既に説明したように、テキストデータＴＸ１が１００ｋバイト程度（又はこの整数倍）であるときに、このテキストデータＴＸ１に対して１／２０程度のデータ量のパリティ情報を得たい場合には、例えばＮの値が６４、Ｍの値が１２８に設定される。以下では説明を簡単にするために、図７に示すようにテキストデータＴＸ１を４行で、かつ３２列に分割した場合を想定する。即ち、Ｎ＝４，Ｍ＝３２とする。この場合、本例では端数は生じないが、例えば図７において、最後の部分テキストデータＴ（４，３２）中の文字が１６個より少ない場合には、足りない部分には予め定めた文字（例えば文字Ａ）をダミーデータとして付加すればよい。また、部分テキストデータＴ（ｉ，ｊ）の長さは、１６文字以外の任意の長さでよいが、処理速度を高めるためには、８文字の倍数が効率的である。

更に、情報処理装置１０は、図７の１６文字分の各部分テキストデータＴ（ｉ，ｊ）をそれぞれ所定の変換テーブルに従って１２８（＝１６×８）ビットのバイナリーデータ（数値データ）よりなる部分データＡ（ｉ，ｊ）に変換する。本例ではその変換テーブルとして、次のように部分テキストデータＴ（ｉ，ｊ）を単にアスキーコードに変換する関数ａｓｃ（Ｔ（ｉ，ｊ））を用いる。

Ａ（ｉ，ｊ）＝ａｓｃ（Ｔ（ｉ，ｊ））　…（１３）

なお、図７にＴ（３，１１）の変換例で示すように、関数ａｓｃ（Ｔ（ｉ，ｊ））は、部分テキストデータＴ（ｉ，ｊ）の先頭の文字のコードが最下位桁となり、末尾の文字のコードが最上位桁となるように変換を行う。この際に、例えば最後の列の部分データＡ（ｉ，３２）が１２８ビットにならないときには、その上位に予め定めた文字コード、又は数値データの０（hex(000･･･)）などのダミーデータが付加される。この結果、図８に示す４行で、３２列の部分データＡ（ｉ，ｊ）が得られる。また、部分データＡ（ｉ，ｊ）を対応するヌクレオチドの配列方向に連続して配列したときの集合体（数値データ）をバイナリーデータＢＮ１とする。図７の部分テキストデータＴ（ｉ，ｊ）と図８の部分データＡ（ｉ，ｊ）とは実質的に同じデータ量である。

次に、本例では、部分データＡ（ｉ，ｊ）をガロア体(Galois field)ＧＦ（２^{m}）上の元のベクトル表示とみなして、部分データＡ（ｉ，ｊ）に対してガロア体ＧＦ（２^{m}）上の所定の演算を施す。本例の部分データＡ（ｉ，ｊ）は１２８ビットであるため、ｍの値は１２８（６４の２倍）となり、ガロア体ＧＦ（２^{128}）が使用される。また、本例ではガロア体ＧＦ（２^{128}）上の既約多項式ＧＦ（Ｘ）及び生成元αとして次の式を使用する。なお、ガロア体ＧＦ（２^{m}）は、拡大ガロア体とも呼ばれることがある。

$$GF(X)=1+X^{121}+X^{126}+X^{127}+X^{128} \quad \cdots(14)$$

$$\alpha=X \quad \cdots(15)$$

ガロア体ＧＦ（２^{128}）上のベクトル表示では、ＧＦ（Ｘ）はbin(1110000100･･･01)となり、αはbin(00･･･0010)となる。なお、生成元αとしては、（１＋Ｘ）なども使用できる。また、ガロア体ＧＦ（２^{128}）上の既約多項式としては、例えば次の既約多項式ＧＦ’（Ｘ）も使用でき、この既約多項式ＧＦ’（Ｘ）に対する生成元としては次のα’などを使用できる。

$$GF'(X)=1+X^{11}+X^{124}+X^{125}+X^{126}+X^{127}+X^{128} \quad \cdots(14A)$$

$$\alpha'=1+X+X^{2} \quad \cdots(15A)$$

また、多項式ＧＦ（Ｘ）が既約かどうかを確かめる最も簡単な方法は、その多項式ＧＦ（Ｘ）の次数をｍとして、ｍ／２を超えない整数をｍ’とすると、その多項式ＧＦ（Ｘ）を次数ｍ’以下の全ての多項式で割ってみることである。このときに、割り切れる多項式がなければ、多項式ＧＦ（Ｘ）は既約である。

また、特に次数ｍの大きい多項式ＧＦ（Ｘ）が既約であることは、例えば文献４（Van der Waerden, B. L. (1953), Modern Algebra(2 vols.), p.77, Ungar, New York）に記載してある「Kroneckerの方法」で確認することができる。また、ＧＦ（Ｘ）が既約であることは、実用的には、ウェブサイト３(http://archives.math.utk.edu/software/msdos/number.theory/ubasic/.html)、又はウェブサイト４(http://www.rkmath.rikkyo.ac.jp/~kida/ubasic.htm)に開示されている整数論研究用のソフトウェアである「UBASIC」中の組み込み関数「POLFACT2」を用いても確認することができる。

また、ガロア体ＧＦ（２^{m}）上の生成元αは、ｋ＝２^{m}－１　とおくと、既約

多項式ＧＦ（Ｘ）を法として、次の関係を満たす。

$\alpha^{k}=1$（ｍｏｄ　ＧＦ（Ｘ））　…（１６）

$\alpha^{k'}\neq 1$（ｍｏｄ　ＧＦ（Ｘ））（$1\leqq k'<k$）　…（１７）

そこで、素数ｐ１，ｐ２，…，ｐｒ及び整数ｎ１，ｎ２，…，ｎｒを用いて、ｋが次のように因数分解できるものとする。

$k=2^{m}-1=p1^{n1}\cdot p2^{n2}\cdot\cdots\cdot pr^{nr}$　…（１８）

このとき、生成元αとは、既約多項式ＧＦ（Ｘ）を法として、αの（$p1^{n1-1}\cdot p2^{n2}\cdot\cdots\cdot pr^{nr}$）乗、（$p1^{n1}\cdot p2^{n2-1}\cdot\cdots\cdot pr^{nr}$）乗、…、（$p1^{n1}\cdot p2^{n2}\cdot\cdots\cdot pr^{nr-1}$）乗が何れも１とならないものであればよい。

また、ガロア体ＧＦ（2^{m}）上の任意の０以外の元βについても（１６）式が成立するため、ｋ（$=2^{m}-1$）を用いてβの逆元β^{-1}は次のように計算することも可能である。

$\beta^{-1}=\beta^{k-1}$　（ｍｏｄ　ＧＦ（Ｘ））　…（１６Ｒ）

従って、例えば部分データＡ（ｉ，ｊ）をβで除算する場合には、部分データＡ（ｉ，ｊ）にβ^{k-1}を乗算すればよい

次のステップ１０６において、情報処理装置１０は、図８の各行（ｉ＝１～４）の部分データＡ（ｉ，ｊ）に対してガロア体ＧＦ（2^{128}）上で、非配列方向（ｊ＝１～３２）に対する和である第１パリティ(Parity)Ｂ１（ｉ）、$\Sigma\alpha^{(j-1)}\cdot A(i,j)$である第２パリティＢ２（ｉ）、及び$\Sigma\alpha^{2(j-1)}\cdot A(i,j)$である第３パリティＢ３（ｉ）を計算する。これらの非配列方向のパリティＢ１（ｉ）～Ｂ３（ｉ）（第１組のパリティ情報）は、生成元αを用いて、かつ既約多項式ＧＦ（Ｘ）を法として以下のように表すことができる。パリティＢ１（ｉ）～Ｂ３（ｉ）における記号（Σ）は係数ｊに対する１～３２の和を意味しており、以下の式は係数ｉの１～４の範囲で計算される。

$B1(i)=\Sigma A(i,j)=A(i,1)+A(i,2)+\cdots+A(i,32)$　…（１９）

$B2(i)=\Sigma\alpha^{(j-1)}\cdot A(i,j)=A(i,1)+\alpha\cdot A(i,2)+\cdots+\alpha^{31}\cdot A(i,32)$　…（２０）

$B3(i)=\Sigma\alpha^{2(j-1)}\cdot A(i,j)=A(i,1)$

$$+\alpha^{2}\cdot A(i,\ 2)+\cdots+\alpha^{62}\cdot A(i,\ 32) \quad \cdots(21)$$

この場合、（１９）式のパリティＢ１（ｉ）のベクトル表示は、部分データＡ（ｉ，ｊ）についてビット毎に排他的論理和演算を行って得られる結果と同じである。また、（２０）式、（２１）式のパリティＢ２（ｉ），Ｂ３（ｉ）は、それぞれ部分データＡ（ｉ，ｊ）を（１）式のように１２７次（ｍ＝１２８）以下の多項式で表して、既約多項式ＧＦ（Ｘ）を法として演算を行うことによって計算することができる。

更に、情報処理装置１０は、図８の各列（ｊ＝１～３２）の部分データＡ（ｉ，ｊ）に対してガロア体ＧＦ（2^{128}）上で、配列方向（ｉ＝１～４）に対する和である第１パリティＣ１（ｊ）、$\Sigma\alpha^{(i-1)}\cdot A(i,\ j)$である第２パリティＣ２（ｊ）、及び$\Sigma\alpha^{2(i-1)}\cdot A(i,\ j)$である第３パリティＣ３（ｊ）を計算する。これらの配列方向のパリティＣ１（ｊ）～Ｃ３（ｊ）（第２組のパリティ情報）は、生成元αを用いて、かつ既約多項式ＧＦ（Ｘ）を法として以下のように表すことができる。パリティＣ１（ｊ）～Ｃ３（ｊ）における記号（Σ）は係数ｉに対する１～４の和を意味しており、以下の式は係数ｊの１～３２の範囲で計算される。

$$C1(j)=\Sigma A(i,\ j)=A(1,\ j)+A(2,\ j)+\cdots+A(4,\ j) \quad \cdots(22)$$

$$C2(j)=\Sigma\alpha^{(i-1)}\cdot A(i,\ j)=A(1,\ j)+\alpha\cdot A(2,\ j)+\cdots+\alpha^{3}\cdot A(4,\ j) \quad \cdots(23)$$

$$C3(j)=\Sigma\alpha^{2(i-1)}\cdot A(i,\ j)=A(1,\ j)+\alpha^{2}\cdot A(2,\ j)+\cdots+\alpha^{6}\cdot A(4,\ j) \quad \cdots(24)$$

部分データＡ（ｉ，ｊ）に対して実際にパリティＢ１（ｉ）～Ｂ３（ｉ）、及びパリティＣ１（ｊ）～Ｃ３（ｊ）を計算した結果のベクトル表示が、図８に１６進数表示で示されている。この例においては、各行のパリティＢ１（ｉ）～Ｂ３（ｉ）、及び各列のパリティＣ１（ｊ）～Ｃ３（ｊ）はそれぞれ３個であるため、２つのヌクレオチドの配列の比較を行う場合に、各行及び各列において、それぞれ３個までの部分データＡ（ｉ，ｊ）の相違部の復元を正確に行うことができる。従って、各行及び各列において、部分データＡ（ｉ，ｊ）の相違部の位置

の検出（特定）だけを行うと共に、相違部の復元を１個だけ行えばよい場合には、パリティ情報として、Ｂ１（ｉ）及びＣ１（ｊ）、又はＢ２（ｉ）及びＣ２（ｊ）だけを使用（計算）するだけでもよい。後者のパリティＢ２（ｉ），Ｃ２（ｊ）だけを使用する場合には、例えば或る行又は列で２つの部分データＡ（ｉ，ｊ）が入れ替わったような配列であっても、配列の相違部の位置を検出できる利点がある。

また、各行及び各列において、それぞれ２個までの部分データＡ（ｉ，ｊ）の相違部の復元を正確に行うことができればよい場合には、第１組のパリティ情報としてＢ１（ｉ），Ｂ２（ｉ），Ｂ３（ｉ）の何れか２つ、及び第２組のパリティ情報としてＣ１（ｊ），Ｃ２（ｊ），Ｃ３（ｊ）の何れか２つだけを使用（計算）するだけでもよい。また、各行と各列とで復元できる部分データの個数が違ってもよい場合には、第１組のパリティ情報と第２組のパリティ情報とでパリティの個数が違ってもよい。更に、各行又は各列において、それぞれ４個以上の相違部の復元を正確に行うためには、例えば$\Sigma \alpha^{s(j-1)} \cdot A(i, j)$であるパリティＢｓ（ｉ）（ｓ＝４，５，…）、又は$\Sigma \alpha^{t(i-1)} \cdot A(i, j)$であるパリティＣｔ（ｊ）（ｔ＝４，５，…）を計算すればよい。

また、図８の部分データＡ（ｉ，ｊ）の配列が実際には、６４行×１２８列であるとすると、図８の例のように、各行及び各列で３個までの相違部の復元を行う場合には、パリティＢ１（ｉ）～Ｂ３（ｉ），Ｃ１（ｊ）～Ｃ３（ｊ）はそれぞれ１２８ビット（１６バイト）であるため、全部のパリティ情報のデータ量は、５７６・１６（＝（６４＋１２８）・３・１６）バイトとなる。一方、部分データＡ（ｉ，ｊ）のデータ量は、８１９２・１６（＝６４・１２８・１６）バイトとなる。従って、全部のパリティ情報のデータ量は、全部の部分データ（ｉ，ｊ）に対してほぼ１／１４程度に減少している。

次に図４のステップ１０７において、情報処理装置１０は、標準試料Ｅの名前の情報、配列の数ＮＡ１、バイナリーデータＢＮ１、パリティＢ１（ｉ）～Ｂ３（ｉ），Ｃ１（ｊ）～Ｃ３（ｊ）を磁気ディスク装置１７のワーキングファイル２０に記録する。この際に、ワーキングファイル２０を複数のファイルとして、バイナリーデータＢＮ１と、パリティＢ１（ｉ）～Ｂ３（ｉ），Ｃ１（ｊ）～Ｃ

３（ｊ）とを別のファイルに記録してもよい。更に、バイナリーデータＢＮ１と共に、ステップ１０２で算出した要約値ＡＢ１をワーキングファイル２０に記録してもよい。

また、バイナリーデータＢＮ１が長いときには、バイナリーデータＢＮ１を複数のファイルに分割して記録してもよい。更に、図７のテキストデータＴＸ１（ひいては図８のバイナリーデータＢＮ１）がかなり長い場合には、テキストデータＴＸ１を例えば数１００ｋバイト程度を単位として複数のデータ群に分割し、各データ群毎にパリティＢ１（ｉ）～Ｂ３（ｉ），Ｃ１（ｊ）～Ｃ３（ｊ）を求めるようにしてもよい。

更に、ステップ１０７において、ＤＮＡ情報の供給者は、ワーキングファイル２０に記録した情報、即ち標準試料Ｅの名前の情報、配列の数ＮＡ１、バイナリーデータＢＮ１、パリティＢ１（ｉ）～Ｂ３（ｉ），Ｃ１（ｊ）～Ｃ３（ｊ）と、マスターファイル１９に記録した要約値ＡＢ１，ＡＢＲ１の情報とを、ＣＤ－Ｒ／ＲＷドライブ１５を介してＣＤ－Ｒ１６に記録してもよい。このＣＤ－Ｒ１６から、更に多数のＣＤ－ＲＯＭを作製してもよく、これらの記録媒体が郵送等によってユーザに販売される。

次の、ステップ１０８において、情報処理装置１０は、標準試料Ｅの名前の情報、配列の数ＮＡ１、配列ＳＴ１，ＳＢ１、要約値ＡＢ１、逆方向の配列ＳＴＲ１，ＳＢＲ１、及び逆方向の要約値ＡＢＲ１を磁気ディスク装置１７のコンテンツファイル２１に記録する。図７のテキストデータＴＸ１が仮に１００Ｍバイト程度の膨大なものであっても、コンテンツファイル２１に記録されるデータは５００バイト程度の僅かなものである。更に、情報処理装置１０は、コンテンツファイル２１中の情報を通信ネットワーク１を介してコンテンツのプロバイダ３に送信する。これによって、コンテンツファイル２１中の情報はプロバイダ３のサーバ内の閲覧可能なコンテンツファイル３１に記録されて、第３者がインターネットを介して自由に閲覧できるようになる。

次のステップ１０９において、ＤＮＡ情報の供給者は、ユーザから購入要求が来るのを待つ状態となる。そして、（ａ）ユーザから標準試料Ｅに対する簡易データの要求があったときには、ステップ１１０に移行して、情報処理装置１０は、

磁気ディスク装置１７のワーキングファイル２０の中のパリティ情報（パリティＢ１（ｉ）～Ｂ３（ｉ），Ｃ１（ｊ）～Ｃ３（ｊ））を例えば電子メールの添付ファイルとしてそのユーザに送信する。一方、ステップ１０９において、（ｂ）ユーザから完全データの要求があったときには、ステップ１１１に移行して、情報処理装置１０は、ワーキングファイル２０中のバイナリーデータＢＮ１をＺＩＰファイル等の形式で圧縮し、この圧縮されたデータを例えば電子メールの添付ファイルとしてそのユーザに送信する。この際に必要に応じて、ハッシュ関数による要約値ＡＢ１を同時に送信してもよい。本例によれば、簡易データ（パリティ情報）はデータ量が少ないために短時間で送信することができる。

また、ステップ１０９において、ユーザは、必要に応じて部分データ、即ち図８の全部の部分データＡ（ｉ，ｊ）の内の所望のデータ、例えば２つの部分データＡ（４，１６）及びＡ（１，１７）のみをその供給者から購入するようにしてもよい。これによって、必要な正確なデータのみを短時間に入手することができる。

次に、ＤＮＡ情報のユーザ（図１のコンピュータシステム２Ｂの所有者とする）側では、図５のステップ１２１において、図１の通信ネットワーク１（インターネット）を介してプロバイダ３のサーバ内のコンテンツファイル３１の内容を閲覧し、その中からステップ１０８で供給者（図１のコンピュータシステム２Ａ）から送信された情報、即ち標準試料Ｅの名前の情報、ヌクレオチドの配列の数ＮＡ１、配列ＳＴ１，ＳＢ１、要約値ＡＢ１、逆方向の配列ＳＴＲ１，ＳＢＲ１、及び逆方向の要約値ＡＢＲ１を読み取り、読み取った情報をコンピュータシステム２Ｂ内の記憶装置の一時ファイルに記録する。

次の、ステップ１２２において、そのユーザは、不図示のＤＮＡのシーケンサーを用いて、標準試料Ｅと同じ種類で検査対象の試料ＦのＤＮＡ中の一方の一列のヌクレオチドの配列を読み取り、読み取られた配列を示すテキストデータＴＸ２（アスキーコードとする）をコンピュータシステム２Ｂ内の情報処理装置に取り込む。その検査対象の試料Ｆとは、例えば突然変異を起こしていると思われる大腸菌であり、そのテキストデータＴＸ２は、標準試料ＥのテキストデータＴＸ１と同様に最初から２０４８個までのヌクレオチドの配列を示すものとする。

試料ＦのＤＮＡ配列は配列番号２に示されている。後述の図９のテキストデータは、配列番号２の配列から数字データを除いて、ａ，ｇ，ｃ，ｔの文字をそれぞれＡ，Ｇ，Ｃ，Ｔで置き換えたものに相当する。

図９は、その試料ＦのＤＮＡのヌクレオチドの配列に対応するテキストデータＴＸ２を示し、この図９の配列の内のアンダーラインを付した部分のみが、図７の標準試料Ｅの配列と異なっている。即ち、試料Ｆの配列は、標準試料Ｅの部分テキストデータＴ（４，１６），Ｔ（１，１７）の部分だけが以下のように異なっている。なお、この段階では、ユーザは、試料Ｆの配列と標準試料Ｅの配列とのどの部分が相違しているのかは分からない。

標準試料Ｅ　　試料Ｆ

T(4, 16)=ATTTGGACGGACGTTG → ATTTGGACATTATGGC

T(1, 17)=ACGGGGTCTATACCTG → GGCCAACTTATACCTG

そして、ユーザのコンピュータシステム２Ｂ側の情報処理装置においても、ＤＮＡの配列情報を処理するためのアプリケーション・プログラムが起動されている。そして、その情報処理装置は、ステップ１２３において、読み取られたテキストデータＴＸ２に上記のＭＤ５ハッシュ関数を施して１２８ビットの要約値ＡＢ２を求めると共に、そのヌクレオチドの配列の数ＮＡ２、及び先頭と末尾との８個ずつのヌクレオチドの配列ＳＴ２，ＳＢ２を求め、これらを内部の記憶装置の第１データファイルに記録する。テキストデータＴＸ２（図９）に対する具体的な値は下記の通りである。

ＡＢ２＝hex(1457b51222a83c3222e87cb4d4e63305) …（２５）

ＮＡ２＝２０４８

ＳＴ２＝ＡＧＣＴＴＴＴＣ，ＳＢ２＝ＣＧＣＧＡＡＧＧ

次のステップ１２４において、情報処理装置は、試料Ｆの配列数ＮＡ２と標準試料Ｅの配列数ＮＡ１とが等しいかどうかを調べ、両者が異なっている場合には、ユーザはステップ１２５に移行して、別のＤＮＡ情報を検索し、ＮＡ２と同じ配列数のＤＮＡ情報をサーチする。本例では、ステップ１２４において、ＮＡ２＝ＮＡ１であるため、動作はステップ１２６に移行して、試料Ｆの先頭と末尾との一部の配列ＳＴ２，ＳＢ２が、標準試料Ｅの配列ＳＴ１，ＳＢ１と等しいかどう

か、更に試料Ｆの要約値ＡＢ２が標準試料Ｅの要約値ＡＢ１（ステップ１２１で一時ファイルに記録されている）と等しいかどうかを調べる。これらが共に等しい場合には、試料Ｆの配列と標準試料Ｅの配列とは非常に高い確率（ほぼ１／２^{128} ≒１／１０^{38}程度の確率）で一致しているとみなすことができる。従って、ステップ１２７に移行して、コンピュータシステム２Ｂの情報処理装置は、その第１データファイルに「試料ＦのＤＮＡ構造は、標準試料ＥのＤＮＡ構造と同一である。」との情報を記録する。

但し、本例では、ＳＴ２＝ＳＴ１，ＳＢ２＝ＳＢ１が成立するが、（１１）式及び（２５）式よりＡＢ２≠ＡＢ１であるため、動作はステップ１２６からステップ１２８に移行して、その情報処理装置は、試料Ｆの先頭と末尾との一部の配列ＳＴ２，ＳＢ２が、標準試料Ｅを逆に並べた配列の一部の配列ＳＴＲ１，ＳＢＲ１と等しいかどうか、更に試料Ｆの要約値ＡＢ２が標準試料Ｅを逆に並べた配列の要約値ＡＢＲ１と等しいかどうかを調べる。これらが共に等しい場合には、試料Ｆの配列と標準試料Ｅを逆に並べた配列とは非常に高い確率で一致しているとみなすことができる。従って、ステップ１３９に移行して、コンピュータシステム２Ｂの情報処理装置は、その第１データファイルに「試料ＦのＤＮＡ構造は、標準試料ＥのＤＮＡ構造に対して回文（palindrome）の関係にある。」との情報を記録する。

本例では、ＳＴ２≠ＳＴＲ２，ＳＢ２≠ＳＢＲ２、かつ（１２）式及び（２５）式よりＡＢ２≠ＡＢＲ１であるため、動作はステップ１２８からステップ１２９に移行して、そのユーザは、通信ネットワーク１（インターネット）を介してＤＮＡ情報の供給者から上記の簡易データ、即ち標準試料Ｅのパリティ情報（Ｂ１（ｉ）～Ｂ３（ｉ），Ｃ１（ｊ）～Ｃ３（ｊ））（図８の情報）を購入し、購入した情報をコンピュータシステム２Ｂ（情報処理装置）内の記憶装置の第２データファイルに記録する。

次に、図６のステップ１３０において、コンピュータシステム２Ｂの情報処理装置は、図９に示すように、試料Ｆのテキストデータ ＴＸ２を配列方向（ヌクレオチドの配列方向）にＮ行で、非配列方向にＭ列の１６文字の長さの部分テキストデータＴＦ（ｉ，ｊ）（ｉ＝１～Ｎ，ｊ＝１～Ｍ）に分割する。分割数Ｎ，Ｍ

は標準試料Ｅの分割数と同じであり、本例では、Ｎ＝４，Ｍ＝３２である。更に、情報処理装置は、図９の各部分テキストデータＴＦ（ｉ，ｊ）を次のようにテキストデータを単にアスキーコードに変換する関数ａｓｃ（ＴＦ（ｉ，ｊ））を用いて、１２８（＝１６・８）ビットのバイナリーデータ（数値データ）よりなる部分データＡＦ（ｉ，ｊ）に変換する。この場合にも、部分テキストデータＴＦ（ｉ，ｊ）の文字列は反転してアスキーコード列に変換される。

ＡＦ（ｉ，ｊ）＝ａｓｃ（ＴＦ（ｉ，ｊ））　…（２６）

この結果、図１０に示す４行で、３２列の部分データＡＦ（ｉ，ｊ）が得られる。また、部分データＡＦ（ｉ，ｊ）を連続して配列した集合体（数値データ）をバイナリーデータＢＮ２とする。

次に、情報処理装置は、ステップ１０６の動作と同様にして、図１０の各行（ｉ＝１～４）の部分データＡＦ（ｉ，ｊ）に対してガロア体ＧＦ（2^{128}）上で、非配列方向（ｊ＝１～３２）に対する和である第１パリティ(Parity)Ｂ１Ｆ（ｉ）、$\Sigma\alpha^{(j-1)}\cdot$ＡＦ（ｉ，ｊ）である第２パリティＢ２Ｆ（ｉ）、及び$\Sigma\alpha^{2(j-1)}\cdot$ＡＦ（ｉ，ｊ）である第３パリティＢ３Ｆ（ｉ）を計算する。これらの非配列方向のパリティＢ１Ｆ（ｉ）～Ｂ３Ｆ（ｉ）（第１組のパリティ情報）は、（１５）式の生成元αを用いて、かつ（１４）式の既約多項式ＧＦ（Ｘ）を法として（１９）式～（２１）式と同様に、係数ｉについて１～４の範囲で計算される。

次に、その情報処理装置は、図１０の各列（ｊ＝１～３２）の部分データＡＦ（ｉ，ｊ）に対してガロア体ＧＦ（2^{128}）上で、配列方向（ｉ＝１～４）に対する和である第１パリティＣ１Ｆ（ｊ）、$\Sigma\alpha^{(i-1)}\cdot$ＡＦ（ｉ，ｊ）である第２パリティＣ２Ｆ（ｊ）、及び$\Sigma\alpha^{2(i-1)}\cdot$ＡＦ（ｉ，ｊ）である第３パリティＣ３Ｆ（ｊ）を計算する。これらの配列方向のパリティＣ１（ｊ）～Ｃ３（ｊ）（第２組のパリティ情報）も、（１５）式の生成元αを用いて、かつ（１４）式の既約多項式ＧＦ（Ｘ）を法として（２２）式～（２４）式と同様に、係数ｊについて１～３２の範囲で計算される。

部分データＡＦ（ｉ，ｊ）に対して実際にパリティＢ１Ｆ（ｉ）～Ｂ３Ｆ（ｉ）、及びパリティＣ１Ｆ（ｊ）～Ｃ３Ｆ（ｊ）を計算した結果のベクトル表

示が、図１０に１６進数表示で示されている。

次に、ステップ１３１において、その情報処理装置は、供給者から購入した簡易データの２組のパリティ、即ち図８（標準試料Ｅ）の２組のパリティＢ１（ｉ）～Ｂ３（ｉ），Ｃ１（ｊ）～Ｃ３（ｊ）と、図１０（試料Ｆ）の２組のパリティＢ１Ｆ（ｉ）～Ｂ３Ｆ（ｉ），Ｃ１Ｆ（ｊ）～Ｃ３Ｆ（ｊ）とを比較して、相違するパリティをサーチする。本例では、図８（標準試料Ｆ）に対して図１０（試料Ｆ）のｉ＝１，４の非配列方向のパリティＢ１Ｆ（１）～Ｂ３Ｆ（１），Ｂ１Ｆ（４）～Ｂ３Ｆ（４）と、ｊ＝１６，１７の配列方向のパリティＣ１Ｆ（１６）～Ｃ３Ｆ（１６），Ｃ１Ｆ（１７）～Ｃ３Ｆ（１７）とが異なっている。なお、各行のパリティＢ１Ｆ（ｉ）～Ｂ３Ｆ（ｉ）、又は各列のパリティＣ１Ｆ（ｊ）～Ｃ３Ｆ（ｊ）において、１つでもパリティが異なっていれば、その行又は列のパリティが異なっているとみなすことができる。

従って、図１０（試料Ｆ）の部分データＡＦ（ｉ，ｊ）において、ｉ＝１，４の行とｊ＝１６，１７の列との交点に位置する４つの部分データＡＦ（１，１６），ＡＦ（４，１６），ＡＦ（１，１７），ＡＦ（４，１７）が図８（標準試料Ｅ）と相違すると特定できる。また、これ以外の試料Ｆの部分データＡＦ（ｉ，ｊ）は標準試料Ｅの部分データＡ（ｉ，ｊ）とほぼ同一であるとみなすことができる。

また、図１１は、主に図１０の試料Ｆのデータ中から図８と異なるパリティＢ１Ｆ（１）～Ｂ３Ｆ（１），Ｂ１Ｆ（４）～Ｂ３Ｆ（４），Ｃ１Ｆ（１６）～Ｃ３Ｆ（１６），Ｃ１Ｆ（１７）～Ｃ３Ｆ（１７）を取り出して表示したものである。また、図１１において図８と異なる部分データＡＦ（１，１６），ＡＦ（４，１６），ＡＦ（１，１７），ＡＦ（４，１７）の位置に、復元すべきデータＸ１，Ｘ２，Ｙ１，Ｙ２を表示している。この復元すべきデータＸ１，Ｘ２，Ｙ１，Ｙ２はそれぞれ図８（標準試料Ｅ）の部分データＡ（１，１６），Ａ（４，１６），Ａ（１，１７），Ａ（４，１７）である。

次のステップ１３２において、その情報処理装置は、図１０の部分データＡＦ（ｉ，ｊ）中で図８の部分データＡ（ｉ，ｊ）と相違する部分データ（ＡＦ（ｉ’，ｊ’）とする）は、各行、又は各列に多くとも３つかどうかを調べる。これ

が成立する場合には、その部分データＡＦ（ｉ’，ｊ’）に対応する標準試料Ｅの部分データＡ（ｉ’，ｊ’）は、ガロア体ＧＦ（2^{128}）上で連立方程式を解くことによって正確に求める（復元する）ことができる。本例では、それが成立する、即ち相違する変換データは、第１行、第４行に２つずつで、かつ第１６列、第１７列に２つずつであるため、動作はステップ１３３に移行する。そして、その情報処理装置は、２組の相違するパリティ、及び試料Ｆの相違する部分データＡＦ（ｉ’，ｊ’）を用いて、図１２のフローチャートに従って対応する標準試料Ｅの部分データＡ（ｉ’，ｊ’）（Ｘ１，Ｘ２，Ｙ１，Ｙ２）を復元する。図１２の計算は、全てガロア体ＧＦ（2^{128}）上で実行される。

この場合、図１１において、第１６列の未知数Ｘ１，Ｘ２は２つであるため、第１６列の２つのパリティＣ１Ｆ（１６），Ｃ２Ｆ（１６）と、対応する図８の２つのパリティＣ１（１６），Ｃ２（１６）と、未知数Ｘ１，Ｘ２に対応する試料Ｆの部分データＡＦ（１，１６），ＡＦ（４，１６）とを用いて２元１次連立方程式を組み立てる。即ち、パリティＣ１（１６），Ｃ１Ｆ（１６）に対する計算式は図１２のステップ１４１の（Ｇ１）式、（Ｇ２）式となる。また、（１５）式の生成元αを用いて、パリティＣ２（１６），Ｃ２Ｆ（１６）に対する計算式はステップ１４２の（Ｇ３）式、（Ｇ４）式となる。

次に、（Ｇ１）式から（Ｇ２）式を引き、（Ｇ３）式から（Ｇ４）式を引くことで、それぞれステップ１４３の（Ｇ５）式、（Ｇ６）式が得られる。（Ｇ５）式、（Ｇ６）式の右辺をそれぞれＣ１Ｘ，Ｃ２Ｘとすることで、２元１次連立方程式が得られる。そこで、これを解くことによって、未知数Ｘ１，Ｘ２はそれぞれステップ１４４の（Ｇ７）式で表すことができる。これを実際に解いた結果、Ｘ１，Ｘ２は次のようになる（図１１参照）。なお、未知数が３個であれば、第３パリティＣ３（１６），Ｃ３Ｆ（１６）なども用いて、３元１次連立方程式を解けばよく、未知数が１個であれば、例えば第１パリティＣ１（１６），Ｃ１Ｆ（１６）などを用いるだけでよい。

Ｘ１＝hex(43475447474347544347544354434154) …（２７）

Ｘ２＝hex(47545447434147474341474754545441) …（２８）

更に、アスキーコード列を文字列に変換する関数ｃｈｒ（）を用いて、この数

値データを文字列に変換すると次のようになる（図１１参照）。この関数ｃｈｒ（）は、上記の関数ａｓｃ（）と対称に、アスキーコード列を１バイト単位で最大桁のコードが末尾の文字となり、最小桁のコードが先頭の文字になるように反転して文字列に変換する。

ｃｈｒ（Ｘ１）＝TACTCTGCTGCGGTGC
＝Ｔ（１，１６）＝ＴＦ（１，１６）　…（２９）

ｃｈｒ（Ｘ２）＝ATTTGGACGGACGTTG＝Ｔ（４，１６）　…（３０）

これより、標準試料Ｅの部分テキストデータＴ（１，１６）と試料Ｆの部分テキストデータＴＦ（１，１６）とは等しく、部分テキストデータＴ（４，１６）だけが部分テキストデータＴＦ（４，１６）（図９参照）と異なることが分かる。

次に、図１１において、第１７列の未知数Ｙ１，Ｙ２についても、第１７列の２つのパリティＣ１Ｆ（１７），Ｃ２Ｆ（１７）と、対応する図８の２つのパリティＣ１（１７），Ｃ２（１７）と、未知数Ｙ１，Ｙ２に対応する試料Ｆの部分データＡＦ（１，１７），ＡＦ（４，１７）とを用いて、図１２のステップ１４５の（Ｇ８）式、（Ｇ９）式よりなる２元１次連立方程式が得られる。これを解くことによって、未知数Ｙ１，Ｙ２はそれぞれステップ１４６の（Ｇ１０）式で表すことができる。これを実際に解いた結果、Ｙ１，Ｙ２は次のようになる（図１１参照）。

Ｙ１＝hex(4754434341544154435447474747434l)　…（３１）

Ｙ２＝hex(4154434354475441474354474141475４)　…（３２）

更に、この数値データ（アスキーコード列）を文字列に変換すると次のようになる（図１１参照）。

ｃｈｒ（Ｙ１）＝ACGGGGTCTATACCTG＝Ｔ（１，１７）　…（３３）

ｃｈｒ（Ｙ２）＝TGAAGTCGATGTCCTA
＝Ｔ（４，１７）＝ＴＦ（４，１７）　…（３４）

これより、標準試料Ｅの部分テキストデータＴ（４，１７）と試料Ｆの部分テキストデータＴＦ（４，１７）とは等しく、部分テキストデータＴ（１，１７）だけが部分テキストデータＴＦ（１，１７）（図９参照）と異なることが分かる。また、本例の方法によって未知数Ｘ１，Ｘ２，Ｙ１，Ｙ２、即ち標準試料Ｅの部

分データＡ（１，１６），Ａ（４，１６），Ａ（１，１７），Ａ（４，１７）が正確に復元できていることが分かる。なお、部分データＡ（１，１６），Ａ（４，１７）は、それぞれ部分データＡＦ（１，１６），ＡＦ（４，１７）と同一であるため、復元されたデータとみなす必要はない。

次のステップ１３４において、その情報処理装置は、復元された部分データＡ（ｉ’，ｊ’）、即ちＡ（４，１６），Ａ（１，１７）で、図１０の試料Ｆのバイナリーデータ ＢＮ２中の対応する部分データＡＦ（４，１６），ＡＦ（１，１７）を置き換えた後、この置き換えによって得られるバイナリーデータＢＮ２をテキストデータＴＸ１’に逆変換する。更に情報処理装置は、そのテキストデータＴＸ１’よりＭＤ５ハッシュ関数を用いて１２８ビットの要約値ＡＢ１’を算出し、この要約値ＡＢ１’が標準試料Ｅの要約値ＡＢ１（ステップ１２１で一時ファイルに記録されている）と等しいかどうかを確認する。本例では、ＡＢ１’＝ＡＢ１が成立するが、例えば図１０の試料Ｆの部分データＡＦ（ｉ，ｊ）中の相違する部分のデータの状態によって、その相違がパリティ情報に反映されないような場合には、その相違する部分の位置がステップ１３２で正確に検出されない可能性がある。このような場合に、ＡＢ１’≠ＡＢ１となったときには、ステップ１３５に移行すればよい。通常は、ＡＢ１’＝ＡＢ１が成立して、動作はステップ１３８に移行して、情報処理装置は、上記の第１データファイルに「試料Ｆの配列と標準試料Ｅの配列との内で相違する部分の位置（ｉ’，ｊ’）、及び相違する部分テキストデータの対」の情報を記録する。本例では、位置（ｉ’，ｊ’）として位置（４，１６），（１，１７）が、相違する部分テキストデータの対としてＡ（４，１６），ＡＦ（４，１６）及びＡ（１，１７），ＡＦ（１，１７）が記録される。

一方、ステップ１３２において、相違する部分データＡＦ（ｉ’，ｊ’）の個数が４個以上の行、又は列が存在する場合には、その行又は列での部分データの正確な復元は困難である。そこで、動作はステップ１３５に移行して、そのユーザはそのＤＮＡ情報の供給者から標準試料Ｅの完全データ、即ち図８のバイナリーデータＢＮ１を通信ネットワーク１（インターネット）を介して購入し、コンピュータシステム２Ｂの情報処理装置は、そのバイナリーデータＢＮ１を記憶装

置の第３データファイルに記録する。

次のステップ１３６において、その情報処理装置は、そのバイナリーデータＢＮ１をテキストデータＴＸ１’に逆変換し、そのテキストデータＴＸ１’よりＭＤ５ハッシュ関数を用いて１２８ビットの要約値ＡＢ１’を算出し、この要約値ＡＢ１’が標準試料Ｅの要約値ＡＢ１（ステップ１２１で一時ファイルに記録されている）と等しいかどうかを確認する。通常は、ＡＢ１’＝ＡＢ１が成立するが、例えば通信エラー等によって送信されたバイナリーデータＢＮ１の中にエラーが生じている場合には、ＡＢ１’≠ＡＢ１となる。このときには、例えば供給者に完全データの再送信を要請する等の対処を行う。そして、ステップ１３６でＡＢ１’＝ＡＢ１が成立するときには、ステップ１３７に移行して情報処理装置は、標準試料Ｅのバイナリーデータ ＢＮ１中で、試料Ｆの相違している部分データＡＦ（ｉ’，ｊ’）に対応する部分データＡ（ｉ’，ｊ’）を求める。その後、動作はステップ１３８に移行する。

なお、上記のステップ１３５では、ユーザはＤＮＡ情報の供給者から完全データ（バイナリーデータＢＮ１）を購入しているが、別の方法として、ステップ１３１で特定された相違する部分データＡＦ（ｉ’，ｊ’）に対応する標準試料Ｅの部分データＡ（ｉ’，ｊ’）のみを購入してもよい。これによって、通信コストを大幅に低減できる。

このように本例のビジネスモデルによれば、第１段階として標準試料Ｅのパリティ情報（Ｂ１（ｉ）～Ｂ３（ｉ），Ｃ１（ｊ）～Ｃ３（ｊ））を購入している。次に、このパリティ情報と試料Ｆのパリティ情報（Ｂ１Ｆ（ｉ）～Ｂ３Ｆ（ｉ），Ｃ１Ｆ（ｊ）～Ｃ３Ｆ（ｊ））とを比較して、相違する部分データＡＦ（ｉ，ｊ）の個数が少ない場合には、対応する標準試料Ｅの部分データＡ（ｉ，ｊ）を復元することとして、相違する部分データの個数が多い場合に、完全データ、又は相違する部分データのみを購入している。従って、初めから膨大な完全データを購入する必要がなく、通信時間を短縮できると共に、情報処理コストを低減できる。

また、本例のパリティ情報を用いれば、ＳＮＰ（一塩基変位多型：Single Nucleotide Polymorphism）のように所定の範囲内で１つのヌクレオチド（塩基）だ

けが異なっているような異常は、容易にその位置の検出、及び復元を行うことができる。

なお、上記の実施の形態では、ＤＮＡ情報のユーザは、ステップ１２１において、コンテンツファイルより標準試料Ｅの配列ＳＴ１，ＳＢ１、要約値ＡＢ１、及び配列ＳＴＲ１，ＳＢＲ１、要約値ＡＢＲ１を読み取って、ステップ１２２～１２８において、標準試料Ｅと試料Ｅとの同一性の判定を行って、両者が異なる場合に標準試料Ｅのパリティ情報（簡易データ）を購入している。しかしながら、標準試料Ｅと試料Ｆとは通常はいくらかは異なっていると考えられるため、このような要約値ＡＢ１等の読み取りから２つの試料の同一性の判定までの動作を省略して、すぐにステップ１２９に移行して、ＤＮＡ情報の供給者から標準試料Ｅのパリティ情報（簡易データ）を購入するようにしてもよい。

なお、上記の実施の形態では、図８に示すように、第１組のパリティ情報（Ｂ１（ｉ）～Ｂ３（ｉ））と第２組のパリティ情報（Ｃ１（ｊ）～Ｃ３（ｊ））とは同じ個数である。しかしながら、図８の例のように部分データＡ（ｉ，ｊ）の配列方向の個数（４個）が非配列方向の個数（３２個）よりも少ない場合には、パリティ情報の全体の情報量を少なくするために、配列方向のパリティ情報として例えばＣ１（ｊ）、Ｃ２（ｊ）、又はＣ１（ｊ）及びＣ２（ｊ）のみを使用するというように、第２組のパリティ情報を第１組のパリティ情報よりも少なくしてもよい。このようにしても、標準試料Ｅと試料Ｆとの相違部は正確に検出できると共に、ＳＮＰのような部分的に生じる相違部は正確に復元することができる。

また、図８（図７）の例のように部分データＡ（ｉ，ｊ）の配列方向の個数が非配列方向の個数よりも少ない場合には、その配列方向をディスプレイの横方向（水平方向）に合わせて表示し、非配列方向にスクロールすることによって、部分データＡ（ｉ，ｊ）及びパリティ情報を作業性の良好な状態で表示することができる。この場合でも、部分データＡ（ｉ，ｊ）をｍビットとすると、その非配列方向の個数は、$(2^m - 1) / 4$　以下であることが望ましい。これによって、非配列方向のパリティ情報として異なる４個のパリティ情報を計算できるため、通常の部分的に生じる相違部は正確に復元することができる。

次に、上記の実施の形態では、ステップ１０５において、図７の標準試料Ｅの

テキストデータＴＸ１を図８の同じデータ量の部分データＡ（ｉ，ｊ）の配列に変換し、この配列からパリティ情報を求めている。その代わりに、データ量を減少させるために、図７の標準試料Ｅのテキストデータ ＴＸ１を表１の変換テーブル（１つのヌクレオチドを２ビットのデータで表すテーブル）を用いてデータ量が１／４のバイナリーデータ（数値データ）に変換し、このバイナリーデータを配列方向、及び非配列方向に分割して、部分データの配列を作成してもよい。

図１３は、そのようにして得られたヌクレオチドの配列方向に５行で、非配列方向に１３列の６４ビット（８バイト）ずつの部分データＢ（ｉ，ｊ）（ｉ＝１～５，ｊ＝１～１３）の配列を１６進数表示で示し、この図１３の各部分データＢ（ｉ，ｊ）は、それぞれ図７の標準試料Ｅ中の３２個のヌクレオチドの配列に対応している。なお、この配列では、最後の部分データＢ（５，１３）に対応する図７の標準試料Ｅは存在しないため、その部分データＢ（５，１３）にはhex(000･･･000)よりなるダミーデータが付加されている。

この場合には、ステップ１０６に対応して、図１３の６４ビットの各部分データＢ（ｉ，ｊ）をガロア体ＧＦ（2^{64}）（ｍ＝６４）上の元のベクトル表示とみなして、部分データＢ（ｉ，ｊ）に対してガロア体ＧＦ（2^{64}）上の所定の演算を施す。ガロア体ＧＦ（2^{64}）上の既約多項式ＧＦ（Ｘ）及び生成元αとしては次の式を使用することができる。

$$GF(X)=1+X^{5}+X^{23}+X^{43}+X^{64} \quad \cdots(35)$$

$$\alpha=X \quad \cdots(36)$$

また、ガロア体ＧＦ（2^{64}）上の既約多項式としては、例えば次の既約多項式ＧＦ’（Ｘ）も使用でき、この既約多項式ＧＦ’（Ｘ）に対する生成元としては次のα’などを使用できる。

$$GF'(X)=1+X^{7}+X^{62}+X^{63}+X^{64} \quad \cdots(35A)$$

$$\alpha'=1+X \quad \cdots(36A)$$

そして、図１の情報処理装置１０は、図１３の各行（ｉ＝１～５）の部分データＢ（ｉ，ｊ）に対してガロア体ＧＦ（2^{64}）上で、非配列方向（ｊ＝１～１３）に対する和である第１パリティ(Parity) Ｂ１Ｂ（ｉ）、$\Sigma\alpha^{(j-1)}\cdot B(i,j)$である第２パリティＢ２Ｂ（ｉ）、及び$\Sigma\alpha^{2(j-1)}\cdot B(i,j)$である第

３パリティＢ３Ｂ（ｉ）を計算する。これらの計算式は、（１９）式～（２１）式に対応しており、この第１組のパリティ情報（Ｂ１Ｂ（ｉ）～Ｂ３Ｂ（ｉ））はそれぞれ６４ビットである。

更に、情報処理装置１０は、図１３の各列（ｊ＝１～１３）の部分データＢ（ｉ，ｊ）に対してガロア体ＧＦ（2^{64}）上で、配列方向（ｉ＝１～５）に対する和である第１パリティＣ１Ｂ（ｊ）、$\Sigma\alpha^{(i-1)}\cdot$Ｂ（ｉ，ｊ）である第２パリティＣ２Ｂ（ｊ）、及び$\Sigma\alpha^{2(i-1)}\cdot$Ｂ（ｉ，ｊ）である第３パリティＣ３Ｂ（ｊ）を計算する。これらの計算式は、（２２）式～（２４）式に対応しており、この第２組のパリティ情報（Ｃ１Ｂ（ｊ）～Ｃ３Ｂ（ｊ））もそれぞれ６４ビットである。

この場合、２組のパリティ情報（Ｂ１Ｂ（ｉ）～Ｂ３Ｂ（ｉ），Ｃ１Ｂ（ｊ）～Ｃ３Ｂ（ｊ））のデータ量は、図８のパリティ情報（Ｂ１（ｉ）～Ｂ３（ｉ），Ｃ１（ｊ）～Ｃ３（ｊ））に比べてほぼ１／４に少なくできる。従って、パリティ情報を通信回線を介して更に短時間で送信することができると共に、記録媒体に記録する場合にも、低容量の記録媒体を使用できる。

この場合には、ユーザ側で図９の試料Ｆのパリティ情報を計算する場合にも、同様にその試料Ｆのテキストデータを表１に従って１／４のデータ量の部分データの配列に変換した後、ガロア体ＧＦ（2^{64}）上の演算によって第１組及び第２組のパリティ情報を計算すればよい。この後の相違するデータの位置の特定、及び元のデータの復元は上記の実施の形態と同様に行うことができる。

なお、上記の実施の形態では、ＤＮＡ又はＲＮＡを構成するヌクレオチドは４種類であるため、テキストデータＴＸ１を少ないデータ量のバイナリーデータに変換する際に、表１に示すように各ヌクレオチドを２ビットのデータで表している。これに対して、ヌクレオチド（又は塩基）を表すテキストデータとして、以下のような１６種類の文字ａ～ｎ（８ビットのアスキーデータ）が使用されることがある。

《表２》

ａ　　アデニン（アデニンを含むヌクレオチドと同義、以下同様）

ｃ　　シトシン

g　グアニン
t　チミン
u　ウラシル
m　アデニン、又はシトシン
r　グアニン、又はアデニン
w　アデニン、又はチミン（若しくはウラシル）
s　グアニン、又はシトシン
y　チミン（若しくはウラシル）、又はシトシン
k　グアニン、又はチミン（若しくはウラシル）
v　アデニン、グアニン、又はシトシン
h　アデニン、シトシン、又はチミン（若しくはウラシル）
d　アデニン、グアニン、又はチミン（若しくはウラシル）
b　グアニン、シトシン、又はチミン（若しくはウラシル）
n　（アデニン、シトシン、グアニン、又はチミン（若しくはウラシル））
又は（不明若しくは他の塩基）。

この場合には、これら１６種類の文字を互いに異なる４ビットのコードに変換し、このコードを用いてテキストデータを数値データ（バイナリーデータ）に変換してもよい。これによって、データ量を１／２にすることができる。また、将来的にヌクレオチド（塩基）の種類が増加したような場合には、これらのヌクレオチドを５ビット、又は６ビットのデータで表現するようにしてもよい。

また、上記の実施の形態では、図７及び図９のヌクレオチドの配列を示すテキストデータよりハッシュ関数によって要約値を算出しているが、情報量としては、それらのテキストデータは例えば表１に従って変換したバイナリーデータ（数値データ）と等価である。従って、これらの変換後のバイナリーデータよりハッシュ関数によってそれぞれ要約値を算出し、これらの算出結果同士を比較するようにしてもよい。そのバイナリーデータのデータ量はテキストデータに対して１／４程度であるため、要約値を算出する時間が短縮できる利点がある。

なお、上記の実施の形態では、ＤＮＡ又はＲＮＡ中のヌクレオチドの配列（又は塩基の配列）の情報を処理対象としているが、本発明は、遺伝子を形成するヌ

クレオチドの配列の情報を処理する場合にも適用できることは言うまでもない。

次に、本発明の実施の形態の他の例につき説明する。本例は、タンパク質又はペプチドを構成するアミノ酸（生物学的物質）の配列情報を処理する場合に本発明を適用したものである。

本例でも基本的に図１のコンピュータシステム２Ａを使用できるが、ＤＮＡのシーケンサー４の代わりに、タンパク質のアミノ酸の配列を決定する配列読み取り装置としてのタンパク質用のシーケンサー（protein Sequencer)が情報処理装置１０に接続される点が異なっている。なお、その配列読み取り装置としては、アミノ酸の配列のデータベースも使用できる。本例でも、例えば新規の試料Ｇのタンパク質のアミノ酸の配列をそのシーケンサーで解読した場合に、その配列を示すテキストデータ（ＴＸ３とする）が情報処理装置１０に供給される。本例では一文字表記を採用するものとして、ｎ個のアミノ酸の配列に対応するテキストデータ（アスキーコードとする）は、ｎバイトの長さである。本例では、その試料Ｇを大腸菌として、そのテキストデータＴＸ３として、図１４に示すように、上記のウェブサイト１から入手した大腸菌の或るタンパク質の８２０個のアミノ酸の配列を示すテキストデータを使用する。

試料Ｇのアミノ酸配列は配列番号３に示されている。図１４のテキストデータは、配列番号３の配列から数字データを除いて、その配列を一文字表記で表したものに相当する。また、図１４においては、そのテキストデータが配列方向（アミノ酸の配列方向）に４行で、その配列方向に直交する非配列方向に２６列の８文字の長さの部分テキストデータＴＧ（ｉ，ｊ）に分割されており、８２１番以上のアミノ酸を示すデータの位置にはダミーデータとして０が付加されている。

次に、情報処理装置１０は、供給されたテキストデータＴＸ３に上記のＭＤ５ハッシュ関数を施して１２８ビットの要約値ＡＢ３を求めると共に、そのアミノ酸の配列の数ＮＡ３、及び先頭と末尾との８個ずつのアミノ酸の配列ＳＴ３，ＳＢ３を求める。テキストデータＴＸ３に対する具体的な値は下記の通りである。

ＡＢ３＝hex(0f66dc2b3024a9739d0e912fde12b8ba) …（４１）

ＮＡ３＝８２０

ＳＴ３＝ＭＲＶＬＫＦＧＧ，ＳＢ３＝ＴＬＳＷＫＬＧＶ

次に、情報処理装置１０は、テキストデータＴＸ３を逆方向に並べ替えたテキストデータＴＸＲ３（＝ＶＧＬＫＷＳ・・・・ＦＫＬＶＲＭ）を求め、このテキストデータＴＸＲ３のＭＤ５ハッシュ関数による要約値ＡＢＲ３、及びこのテキストデータＴＸＲ３の先頭と末尾との８個ずつのアミノ酸の配列ＳＴＲ３，ＳＢＲ３を求める。配列ＳＴＲ３，ＳＢＲ３は、上記の配列ＳＢ３，ＳＴ３をそれぞれ逆方向に並べ替えることによって容易に求めることができる。これらの具体的な値は以下の通りである。テキストデータＴＸＲ３の配列は、テキストデータＴＸ３の配列に対して回文（palindrome）の関係にあると言うことができる。

ＡＢＲ３＝hex(e895f433e1e77f84b3cadeead1a52380) …（４２）

ＳＴＲ３＝ＶＧＬＫＷＳＬＴ，ＳＢＲ３＝ＧＧＦＫＬＶＲＭ

次に、情報処理装置１０は、試料Ｇの名前の情報（試料を特定する情報）、配列の数ＮＡ３、テキストデータＴＸ３、配列ＳＴ３，ＳＢ３、要約値ＡＢ３、逆方向の配列ＳＴＲ３，ＳＢＲ３、及び逆方向の要約値ＡＢＲ３を磁気ディスク装置１７のマスターファイル１９に記録する。この際に、マスターファイル１９を複数のファイルとして、テキストデータＴＸ３と、それ以外のデータとを別のファイルに記録してもよい。続いて、情報処理装置１０は、例えば図７と同様に図１４に示すように、試料Ｇのテキストデータ ＴＸ３を配列方向（アミノ酸の配列方向）にＮ行で、その配列方向に直交する非配列方向にＭ列の８文字（６４ビット）の長さの部分テキストデータＴＧ（ｉ，ｊ）に分割する。Ｎ，Ｍはそれぞれ２以上の任意の整数である。本例ではテキストデータＴＸ３に例えば１２文字分のダミーデータ（本例では０であるが、それ以外に例えば文字Ａなども使用できる）を付加して得られる８３２（＝８・４・２６）バイトのテキストデータ（これをＴＸ３’と呼ぶ）を作成し、テキストデータＴＸ３’をＮ＝４，Ｍ＝２６で分割する。本例では、その８文字分の部分テキストデータＴＧ（ｉ，ｊ）を、次のようにテキストデータを単にアスキーコード（数値データ）に変換する関数ａｓｃ（）を用いて、そのまま６４ビットの部分データＡＧ（ｉ，ｊ）として扱う。

ＡＧ（ｉ，ｊ）＝ａｓｃ（ＴＧ（ｉ，ｊ））　…（４３）

なお、図１４にＴＧ（３，１１）の変換例で示すように、関数ａｓｃ（ＴＧ（ｉ，ｊ））は、部分テキストデータＴＧ（ｉ，ｊ）の先頭の文字のコードが最

下位桁となり、末尾の文字のコードが最上位桁となるように反転して変換を行う。なお、この際に、各アミノ酸を６ビットのデータで表してもよいが、データ量は３／４程度になるだけであるため、本例では部分テキストデータ（アスキーコード列）をそのまま部分データ（数値データ）として扱う。

図１５は、試料Ｇの部分データＡＧ（ｉ，ｊ）の配列を示している。それに続いて図１３の例と同様に、情報処理装置１０は、その図１５の４行で２６列の６４ビットの部分データＡＧ（ｉ，ｊ）をガロア体ＧＦ（2^{64}）（ｍ＝６４）上の元のベクトル表示とみなして、部分データＡＧ（ｉ，ｊ）に対してガロア体ＧＦ（2^{64}）上の所定の演算を施す。ガロア体ＧＦ（2^{64}）上の既約多項式ＧＦ（Ｘ）及び生成元αとしては、（３５）式（又は（３５Ａ）式）、及び（３６）式（又は（３６Ａ）式）などを使用できる。

具体的に、情報処理装置１０は、図１５の各行（ｉ＝１～４）の部分データＡＧ（ｉ，ｊ）に対してガロア体ＧＦ（2^{64}）上で、非配列方向（ｊ＝１～２６）に対する和である第１パリティ(Parity)Ｂ１Ｇ（ｉ）、$\Sigma\alpha^{(j-1)}\cdot$ＡＧ（ｉ，ｊ）である第２パリティＢ２Ｇ（ｉ）、及び$\Sigma\alpha^{2(j-1)}\cdot$ＡＧ（ｉ，ｊ）である第３パリティＢ３Ｇ（ｉ）を計算する。これらの計算式は、（１９）式～（２１）式に対応しており、この第１組のパリティ情報（Ｂ１Ｇ（ｉ）～Ｂ３Ｇ（ｉ））はそれぞれ６４ビットである。

更に、情報処理装置１０は、図１５の各列（ｊ＝１～２６）の部分データＡＧ（ｉ，ｊ）に対してガロア体ＧＦ（2^{64}）上で、配列方向（ｉ＝１～４）に対する和である第１パリティＣ１Ｇ（ｊ）、$\Sigma\alpha^{(i-1)}\cdot$ＡＧ（ｉ，ｊ）である第２パリティＣ２Ｇ（ｊ）、及び$\Sigma\alpha^{2(i-1)}\cdot$ＡＧ（ｉ，ｊ）である第３パリティＣ３Ｇ（ｊ）を計算する。これらの計算式は、（２２）式～（２４）式に対応しており、この第２組のパリティ情報（Ｃ１Ｇ（ｊ）～Ｃ３Ｇ（ｊ））もそれぞれ６４ビットである。このように計算して得られたパリティＢ１Ｇ（ｉ）～Ｂ３Ｇ（ｉ），Ｃ１Ｇ（ｊ）～Ｃ３Ｇ（ｊ）が、図１５に１６進数表示で示されている。

この例においては、各行、各列で３つのパリティを計算しているが、アミノ酸の配列はヌクレオチドの配列に比べるとデータ量がかなり少ないため、実用的には、パリティ情報は各行（非配列方向）、及び各列（配列方向）においてそれぞ

れ１つ（例えばＢ２Ｇ（ｉ）とＣ２Ｇ（ｊ））としてもよい。図１５の例において、パリティＢ２Ｇ（ｉ），Ｃ２Ｇ（ｊ）のみを使用するものとすると、パリティはそれぞれ６４ビット（８バイト）であるため、全部のパリティのデータ量は、２４０（＝８・３０）バイトとなる。従って、全部のパリティのデータ量は、全体の元のテキストデータＴＸ３（８２０バイト）に対してほぼ１／３に減少している。

次に、情報処理装置１０は、試料Ｇの名前の情報、配列の数ＮＡ３、テキストデータＴＸ３、要約値ＡＢ３，ＡＢＲ３、及びパリティ情報を磁気ディスク装置１７のワーキングファイル２０に記録する。この際に、ワーキングファイル２０を複数のファイルとしてもよい。その後、情報処理装置１０は、試料Ｇの名前の情報、配列の数ＮＡ３、配列ＳＴ３，ＳＢ３、要約値ＡＢ３、逆方向の配列ＳＴＲ３，ＳＢＲ３、及び逆方向の要約値ＡＢＲ３を磁気ディスク装置１７のコンテンツファイル２１に記録する。更に、情報処理装置１０は、コンテンツファイル２１中の情報を通信ネットワーク１を介してコンテンツのプロバイダ３に送信する。これによって、コンテンツファイル２１中の情報はプロバイダ３のサーバ内の閲覧可能なコンテンツファイル３１に記録されて、第３者がインターネットを介して自由に閲覧できるようになる。この結果、第３者は、公開されている試料Ｇの配列の数ＮＡ３、及び要約値ＡＢ３（又は必要に応じてＡＢＲ３）を自分の保有するアミノ酸の配列の配列数、及び要約値と比較することによって、その試料Ｇが自分にとって新規かどうかを判定できる。また、ユーザは、その試料Ｇの配列情報を複数の供給者から誤って重複して購入することを回避することができる。

その後、コンピュータシステム２Ａの所有者（アミノ酸情報の供給者）は、ユーザから購入要求が来るのを待つ状態となる。そして、ユーザから試料Ｇに対する簡易データの要求があったときには、情報処理装置１０は、磁気ディスク装置１７のワーキングファイル２０の中の試料Ｇのパリティ情報（例えばその中のＢ２Ｇ（ｉ），Ｃ２Ｇ（ｊ））を例えば電子メールの添付ファイルとしてそのユーザに送信する。パリティの情報を購入したユーザは、試料Ｇと同じ種類の自分で解読した試料のアミノ酸の配列のパリティと、その購入したパリティとを比較す

ることによって、相違する部分の検出及び復元を或る程度行うことができる。

一方、ユーザから完全データの要求があったときには、情報処理装置１０は、ワーキングファイル２０中のテキストデータＴＸ３をＺＩＰファイル等の形式で圧縮し、この圧縮されたデータを例えば電子メールの添付ファイルとしてそのユーザに送信する。この際に必要に応じて、ハッシュ関数による要約値ＡＢ３を同時に送信してもよい。本例によれば、簡易データ（パリティ情報）はデータ量を少なくできるために短時間で送信することができる。

更に、そのアミノ酸の配列情報の供給者は、ワーキングファイル２０に記録した情報、即ち試料Ｇの名前の情報、配列の数ＮＡ３、テキストデータＴＸ３、要約値ＡＢ３，ＡＢＲ３、及びパリティ情報をＣＤ－Ｒ／ＲＷドライブ１５を介してＣＤ－Ｒ１６に記録してもよい。このＣＤ－Ｒ１６から、更に多数のＣＤ－ＲＯＭを作製してもよく、これらの記録媒体が郵送等によってユーザに販売される。

なお、上記の実施の形態では、生物学的物質の配列に対応するテキストデータ、又は該テキストデータを所定の規則に従って変換して得られる数値データを、前記生物学的物質の配列方向に複数行で、かつ前記配列方向に交差する非配列方向に複数列の長さがｍビット（ｍは１６以上の整数）の部分データに分割し、複数行の前記部分データに各行毎に前記非配列方向にガロア体ＧＦ（2^m）上の第１の演算を施して第１組のパリティ情報を求めると共に、複数列の前記部分データに各列毎に前記配列方向にガロア体ＧＦ（2^m）上の第２の演算を施して第２組のパリティ情報を求めている。

これに関して、ｍが１６より小さい場合には、各部分データは、例えば１～７個程度のヌクレオチド、又は１～２個程度のアミノ酸の配列に対応するため、１つの生物学的物質の配列に対して計算すべきパリティ情報の数が多くなり過ぎて好ましくない。また、最近のコンピュータの計算能力を活かしきれないという不都合もある。特に、コンピュータの処理単位が６４ビットの倍数である場合には、パリティ情報の計算効率を高めるために、そのｍの値も６４，１２８，１９２，２５６などの６４の倍数とすることが望ましい。

また、例えば図１３の実施の形態では、ｍビットを超える素数をＰ（＞2^m）とすると、この素数Ｐを法とするガロア体ＧＦ（Ｐ）を用いてパリティ情報を計

算することも可能である。しかしながら、このガロア体ＧＦ（Ｐ）を用いた場合には、ｍビットの部分データに演算を施して得られる個々のパリティ情報がｍビットを超える場合があるため、パリティ情報がガロア対ＧＦ（2^m）を用いた場合に比べて（ｍ＋１）／ｍ倍程度に長くなるという不都合がある。即ち、ガロア体ＧＦ（Ｐ）には演算が容易であるという利点があるが、ガロア体ＧＦ（2^m）を用いた場合には、個々のパリティ情報をｍビットにできるため、パリティ情報を簡潔に記録できる利点がある。

2^m（ｍ≧１６）を超える、又は2^mより僅かに小さい程度の大きい値の素数Ｐを求めるには、例えば ミラー・ラビン（Miller-Rabin）の素数判定法（例えばM.O.Rabin: "Probabilistic algorithms for testing primality", Journal of Number Theory, 12, pp.128-138(1980)参照）を使用することができる。また、生成元を求めるために大きい数の因数分解を行うためには、例えば２次ふるい法(quadratic sieve)（例えばC.Pomerance: "Factoring", In Cryptology and Computational Number Theory, pp.27-47, American Mathematical Society (1990)参照）を使用することができる。また、実際に大きい数の素数Ｐの決定、及び大きい数の因数分解を行うためには、上記の「UBASIC」中の組み込み関数「ECM」を使用してもよい。

そして、素数Ｐを用いるガロア体ＧＦ（Ｐ）は、（０，１，…，Ｐ－１）のＰ個の元より構成され、加減乗除の演算はＰを法として実行される。これを上記の実施の形態に適用する場合には、ｍビットの部分データ（Ａ（ｉ，ｊ）等）をそのガロア体ＧＦ（Ｐ）の何れかの元に対応させればよい。そして、パリティ情報を容易に算出するためには、ガロア体ＧＦ（Ｐ）の生成元δを使用するとよい。ガロア体ＧＦ（Ｐ）上の生成元δは、ｋ＝Ｐ－１　とおくと、Ｐを法として、次の関係を満たす。

$\delta^k = 1 \pmod P$　　…（Ａ２）

$\delta^{k'} \neq 1 \pmod P$　（$1 \leqq k' < k$）　…（Ａ３）

そこで、素数ｐ１，ｐ２，…，ｐｒ及び整数ｎ１，ｎ２，…，ｎｒを用いて、ｋが次のように因数分解できるものとする。

$k = P - 1 = p1^{n1} \cdot p2^{n2} \cdots \cdot pr^{nr}$　…（Ａ４）

このとき、生成元δとは、Ｐを法として、δの（ｐ１$^{n1-1}$・ｐ２^{n2}・…・ｐｒ^{nr}）乗、（ｐ１^{n1}・ｐ２$^{n2-1}$・…・ｐｒ^{nr}）乗、…、（ｐ１^{n1}・ｐ２^{n2}・…・ｐｒ$^{nr-1}$）乗が何れも１とならないものであればよい。

また、ガロア体ＧＦ（Ｐ）上の任意の０以外の元εについても（Ａ２）式が成立するため、ｋ（＝Ｐ－１）を用いてεの逆元ε^{-1}は次のように計算することも可能である。

$$\varepsilon^{-1}=\varepsilon^{k-1} \quad (\mathrm{mod}\ P) \qquad \cdots(A5)$$

従って、例えば或る部分データをεで除算する場合には、その部分データにε^{k-1}を乗算すればよい

そして、ガロア体ＧＦ（Ｐ）を用いる場合にも、図３～図６のＤＮＡ情報の供給者の動作、及びＤＮＡ情報のユーザの動作は、ガロアＧＦ（２^{m}）上の演算をガロア体ＧＦ（Ｐ）上の演算に変えるだけで全て同様に適用することができる。

具体的に図１３（例えば標準試料とする。）の実施の形態において、ガロア体ＧＦ（Ｐ）上でパリティ情報を計算する場合には、部分データＢ（ｉ，ｊ）は６４ビットの全ての値を取る可能性があるため、素数Ｐとしては２^{64}（ｍ＝６４）を超える範囲で６５（＝ｍ＋１）ビット以下の素数（例えば最も小さい素数）を選択すればよい。このような素数Ｐは１６進数表示で次のようになる。また、この素数Ｐに対する生成元δとしては、例えば次の値を使用できる。なお、この場合、例えば３も生成元であるため、生成元として２のべき乗以外の数を使用したいときには、生成元δとして例えば３を使用してもよい。

$$P=\mathrm{hex}(10000000000000000d) \qquad \cdots(A6)$$

$$\delta=2 \qquad \cdots(A7)$$

この条件で、図３のステップ１０６に対応して、図１３の６４ビットの各部分データＢ（ｉ，ｊ）をガロア体ＧＦ（Ｐ）上の元のとみなして、部分データＢ（ｉ，ｊ）に対して生成元δを用いてガロア体ＧＦ（Ｐ）上の所定の演算を施す。一例として、図１３の各行（ｉ＝１～５）の部分データＢ（ｉ，ｊ）に対してガロア体ＧＦ（Ｐ）上で、非配列方向（ｊ＝１～１３）に対する和である第１パリティＢ１Ｂ（ｉ）’、Σδ$^{(j-1)}$・Ｂ（ｉ，ｊ）である第２パリティＢ２Ｂ（ｉ）’、及びΣδ$^{2(j-1)}$・Ｂ（ｉ，ｊ）である第３パリティＢ３Ｂ（ｉ）’を計算す

る。これらの計算式は、（１９）式～（２１）式に対応しており、この第１組のパリティ情報（Ｂ１Ｂ（ｉ）’～Ｂ３Ｂ（ｉ）’）はそれぞれ最大で（６４＋１）ビットである。

更に、図１３の各列（ｊ＝１～１３）の部分データＢ（ｉ，ｊ）に対してガロア体ＧＦ（Ｐ）上で、配列方向（ｉ＝１～５）に対する和である第１パリティＣ１Ｂ（ｊ）’、$\Sigma\delta^{(i-1)}\cdot$Ｂ（ｉ，ｊ）である第２パリティＣ２Ｂ（ｊ）’、及び$\Sigma\delta^{2(i-1)}\cdot$Ｂ（ｉ，ｊ）である第３パリティＣ３Ｂ（ｊ）’を計算する。これらの計算式は、（２２）式～（２４）式に対応しており、この第２組のパリティ情報（Ｃ１Ｂ（ｊ）’～Ｃ３Ｂ（ｊ）’）もそれぞれ最大で（６４＋１）ビットである。

同様に図６のステップ１３０に対応して、比較対象の試料の部分データの配列に対してもガロア体ＧＦ（Ｐ）上で非配列方向の第１組のパリティ情報、及び配列方向の第２組のパリティ情報を計算する。そして、本例でも図６のステップ１３１～１３８に対応して、２つの試料のパリティ情報の比較、相違する部分データの特定、及び相違する部分データの復元を行うことができる。

なお、部分データが１２８ビット（ｍ＝１２８）で全ての値を取る場合に、ガロア体ＧＦ（Ｐ）上でパリティ情報を計算する場合には、2^{128}を超えて１２９ビット以下の素数Ｐ、及びこれに対応する生成元δの一例としては、次の値を使用することができる。

$P=2^{128}+51$　　…（Ａ８）

$\delta=2$　　…（Ａ９）

次に、図７、図８の実施の形態のように部分データＡ（ｉ，ｊ）がテキストデータそのものである場合には、ｍビットの部分データＡ（ｉ，ｊ）の最大値Ｎｍａｘは（2^m-1）よりも小さくなる。具体的に図７のテキストデータとしてＡＳＣＩＩコードを使用する場合、アルファベットのＡＳＣＩＩコードはhex(41)～hex(7a)であるため、最大値Ｎｍａｘは次のようになる。

$Nmax=hex(7a7a7a7a\cdots\cdots7a7a7a)<2^m-1$　…（Ａ１０）

この場合には、ガロア体ＧＦ（Ｐ）を構成するための素数Ｐを次のように、部分データＡ（ｉ，ｊ）の最大値Ｎｍａｘより大きく、且つｍビットの数として選

択することができる。

$2^m > P > Nmax$　　…（Ａ１１）

具体的に、ｍ＝１２８で（Ａ１０）式が成立する場合には、（Ａ１１）式を満たす範囲で最も小さい素数Ｎ及びこれに対する生成元δの一例は次のようになる。

Ｐ＝hex(7a7a7a7a7a7a7a7a7a7a7a7a7a7a7f)　　…（Ａ１２）

δ＝５　　…（Ａ１３）

この素数Ｎ及び生成元δを用いるガロア体ＧＦ（Ｐ）上で図７、図８の部分データＡ（ｉ，ｊ）に対して第１組のパリティ情報、及び第２組のパリティ情報を計算すると、得られる各パリティはそれぞれｍビットである。従って、演算が簡単であると共に、ガロア体ＧＦ（2^m）を用いたときと同じ長さの簡潔なパリティ情報を得ることができる。

同様に、図１４、図１５の実施の形態においても、テキストデータとしてＡＳＣＩＩコードを使用する場合、ｍビット（図１４ではｍ＝６４）の部分データＡＧ（ｉ，ｊ）の最大値Ｎｍａｘは次のようになる。

$Nmax = hex(7a7a7a7a7a7a7a7a) < 2^m - 1$　　…（Ａ１４）

この場合にも、ガロア体ＧＦ（Ｐ）を構成するための素数Ｐを（Ａ１１）式が成立するように選択することができる。

具体的に、ｍ＝６４で（Ａ１４）式が成立する場合には、（Ａ１１）式を満たす範囲で最も小さい素数Ｎ及びこれに対する生成元δの一例は次のようになる。なお、この場合、生成元δとしては例えば５も使用できるため、２の代わりに５などの２のべき乗以外の数を使用するようにしてもよい。

Ｐ＝hex(7a7a7a7a7a7a7ad5)　　…（Ａ１５）

δ＝２　　…（Ａ１６）

また、パリティ情報をガロア体ＧＦ（Ｐ）上で計算する場合にも、部分データの配列方向の個数を、その部分データの非配列方向の個数よりも少なくして、配列方向（第２組）のパリティ情報の個数を、非配列方向（第１組）のパリティ情報の個数よりも少なくしてもよい。これによって、パリティ情報をディスプレイに表示し易いと共に、パリティ情報の情報量が必要以上に大きくなるのを避けることができる。この場合、生成元δを用いて非配列方向で４個までの相違部を正

確に復元するためには、非配列方向の部分データの個数を（Ｐ－１）／４以下にすればよい。

なお、上記の実施の形態では、各部分データに所定の係数を乗じてパリティ情報を求めているが、パリティ情報としては、例えばＢＣＨ符号（Bose-Chaudhari-Hocquenghem Codes）（例えば J.L.Massey: "Shift Register Synthesis and BCH Decoding", IEEE Trans., IT-15, pp.122-127(1969) 参照）を用いてもよい。但し、従来のＢＣＨ符号は小さい値m'（m≦１６程度）のガロア体ＧＦ（$2^{m'}$）で考えられているため、これを本発明で使用されている大きい値m（m≧１６）のガロア体ＧＦ（2^{m}）上で構築し直す必要がある。

ここで、上記の実施の形態で使用するハッシュ関数に関して説明する。通常の暗号理論で使用されるハッシュ関数は、テキストデータ中のスペースコード及び改行コード等も全て演算処理対象としているが、ヌクレオチド及びアミノ酸の配列情報については見やすくするために、例えば配列番号１～３で示すように、途中にスペースコード、順序を示す数字コード、及び改行コードを挿入する場合がある。そこで、ヌクレオチドやアミノ酸などの生物学的物質の配列情報を演算処理対象とするハッシュ関数においては、テキストデータ中の所定コードとしてのスペースコード、数字コード、及び改行コードを無視する機能を付加することが望ましい。また、隣接する文字を”－”（ハイフン）で分けることも考えられるが、この場合には、更に”－”記号も無視する必要がある。また、例えばファイルの最後に「データの終わりを示すコード」が付加されるような場合には、そのコードも無視するようにしてもよい。

また、例えばヌクレオチドの配列が、通常は小文字で表されるような場合には、ハッシュ関数に、選択的に大文字を小文字に変換して要約値を計算する機能を持たせるようにしてもよい。逆に、例えばアミノ酸の配列が、通常は大文字で表されるような場合には、ハッシュ関数に、選択的に小文字を大文字に変換して要約値を計算する機能を持たせるようにしてもよい。

更に、原ファイルを複数の分割ファイルに分割する際には、複数の分割ファイルの順序等を示すデータ（以下、「コメントデータ」と言う）を各分割ファイルに付加することが望ましいことがある。このように分割ファイル、又は１つの原

ファイルにコメントデータを付加する場合にも、コメントデータはハッシュ関数で無視する必要がある。そのため、例えばコメントデータは所定の開始記号（例えば　／＊　）及び終了記号（例えば　＊／）の間に記録し、ハッシュ関数で処理する際に開始記号から終了記号までのデータは無視するようにすればよい。

また、上記の実施の形態では、例えば生物のＤＮＡのヌクレオチドの配列（又はタンパク質のアミノ酸の配列）内の先頭の一部、及び末尾の一部の配列、並びにその配列のテキストデータの要約値をインターネット上で公開することがある。この場合には、その公開されている一部の配列と、その要約値とからそのテキストデータの内容が推定される可能性もある。これを回避するために、そのテキストデータをハッシュ関数で処理する際に、その公開されている配列を除いた部分についてのみ、そのハッシュ関数を施して要約値を求めるようにしてもよい。

なお、上記の実施形態の図５のステップ１２４において、標準試料Ｅの配列の数ＮＡ１と試料Ｆの配列の数ＮＡ２とがｋ（ｋは１以上の整数）だけ異なる場合で、例えばＮＡ２＝ＮＡ１＋ｋのときには、試料Ｆのテキストデータ ＴＸ２からｋ個のヌクレオチド分のデータを取り除いてから、相違部を求めるためにステップ１２６に移行してもよい。また、ＮＡ２＝ＮＡ１－ｋのときには、そのテキストデータＴＸ２にｋ個のヌクレオチド分のダミーデータを付加してから、ステップ１２６に移行してもよい。これによって、試料Ｆの配列の過剰部又は欠損部の位置を特定することができる。このときに、ｋ＝１であるときには、テキストデータＴＸ２中でデータを除去又は付加する位置は、全長の１／２の位置、その前部又は後部の１／２（全体の１／４）の位置、更にその全部又は後部の１／２（全体の１／８）の位置、…としていくことで、最も短時間に過剰部又は欠損部の位置を特定できる。

なお、本発明は上述の実施の形態に限定されず、本発明の要旨を逸脱しない範囲で種々の構成を取り得ることは勿論である。

産業上の利用の可能性

本発明によれば、核酸や遺伝子中のヌクレオチド、又は又はタンパク質やペプチド中のアミノ酸などの生物学的物質の配列情報を、それらの配列を示すテキス

トデータよりも少ないデータ量のパリティ情報として近似的に記録することができる。従って、そのパリティ情報は、低容量の記録媒体にも記録できると共に、通信回線を介して短時間に送信することが可能となる。また、ガロア体ＧＦ（2^m）上の演算を行うことによって、個々のパリティ情報を部分データと同じｍビットの情報量で簡潔に記録できる利点がある。

一方、そのパリティ情報を求める場合に、素数Ｐのガロア体ＧＦ（Ｐ）上の演算を行うことによって、パリティ情報の情報量は（ｍ＋１）／ｍ倍程度に多くなるが、演算を単純化することができる。

特に、部分データの最大値Ｎｍａｘが（2^m －１）よりも小さい値であり、その素数Ｐを、（2^m ＞Ｐ＞Ｎｍａｘ）の関係を満たすように選択できる場合には、演算を単純化した上で、各パリティ情報を部分データと同じｍビットの情報量で簡潔に記録できる利点がある。

また、２つの生物学的物質の配列のパリティ情報を比較することによって、２つの配列間の相違部の位置を少ないデータ量で容易に特定（検出）できると共に、必要に応じてその相違部の情報を復元することができる。従って、例えばＳＮＰ（一塩基変位多型：Single Nucleotide Polymorphism）を少ないデータ量で容易に発見することができる。

また、本発明によれば、ヌクレオチド又はアミノ酸などの生物学的物質の配列情報を近似する情報（パリティ情報）を少ないデータ量でユーザに供給できるビジネスモデルを提供することができる。この場合に、更に数学的な要約値を用いることによって、ユーザが提供された配列情報と情報供給者が保持している配列情報との同一性の確認などを容易に行うことができる。また、同一の複数の配列情報を誤って購入することも防止できる。

請　求　の　範　囲

１．生物学的物質の配列情報の記録方法であって、

前記生物学的物質の配列に対応するテキストデータ、又は該テキストデータを所定の規則に従って変換して得られる数値データを、前記生物学的物質の配列方向に複数行で、かつ前記配列方向に交差する非配列方向に複数列の長さがｍビット（ｍは１６以上の整数）の部分データに分割し、

複数行の前記部分データに各行毎に前記非配列方向にガロア体ＧＦ（2^m）上の第１の演算を施して第１組のパリティ情報を求めると共に、

複数列の前記部分データに各列毎に前記配列方向にガロア体ＧＦ（2^m）上の第２の演算を施して第２組のパリティ情報を求め、

前記第１組及び第２組のパリティ情報で前記生物学的物質の配列を表すことを特徴とする生物学的物質の配列情報の記録方法。

２．請求の範囲１に記載の記録方法であって、

前記ガロア体ＧＦ（2^m）上の生成元をαとしたとき、

前記第１組のパリティ情報は、複数行の各行の前記部分データにそれぞれ前記非配列方向に順次α^{sp}，$\alpha^{s(p+1)}$，$\alpha^{s(p+2)}$，…，$\alpha^{s(p+dp)}$（ｓは０以上の整数、ｐは０以上の整数、ｄｐは１以上の整数）を乗算した後、

該演算で得られた複数の積について各行毎に求められた和を含み、

前記第２組のパリティ情報は、複数列の各列の前記部分データにそれぞれ前記配列方向に順次α^{tq}，$\alpha^{t(q+1)}$，$\alpha^{t(q+2)}$，…，$\alpha^{t(q+dq)}$（ｔは０以上の整数、ｑは０以上の整数、ｄｑは１以上の整数）を乗算した後、

該演算で得られた複数の積について各列毎に求められた和を含むことを特徴とする生物学的物質の配列情報の記録方法。

３．請求の範囲２に記載の記録方法であって、

前記部分データの前記配列方向の個数は、前記部分データの前記非配列方向の個数よりも少なく、

前記第２組のパリティ情報の個数は、前記第１組のパリティ情報の個数よりも少ないことを特徴とする生物学的物質の配列情報の記録方法。

４．請求の範囲３に記載の記録方法であって、

前記部分データの前記非配列方向の個数は、（2^m －１）／４　以下であることを特徴とする生物学的物質の配列情報の記録方法。

５．請求の範囲２に記載の記録方法であって、

前記整数ｓ及びｔは０であることを特徴とする生物学的物質の配列情報の記録方法。

６．請求の範囲２に記載の記録方法であって、

前記整数ｓ及びｔは１であることを特徴とする生物学的物質の配列情報の記録方法。

７．請求の範囲２に記載の記録方法であって、

前記第１組のパリティ情報は、前記複数行の各行毎に前記整数ｓについて互いに異なる複数の値で求めた複数の和を含み、

前記第２組のパリティ情報は、前記複数列の各列毎に前記整数ｔについて互いに異なる複数の値で求めた複数の和を含むことを特徴とする生物学的物質の配列情報の記録方法。

８．請求の範囲１に記載の記録方法であって、

前記生物学的物質の配列中の各生物学的物質をそれぞれ６ビット以下の数値データで表して得られる数値データを前記演算の対象とすることを特徴とする生物学的物質の配列情報の記録方法。

９．請求の範囲１に記載の記録方法であって、

前記ガロア体ＧＦ（2^m）を規定する整数ｍは６４の倍数であることを特徴とする生物学的物質の配列情報の記録方法。

１０．請求の範囲１に記載の記録方法であって、

前記生物学的物質の配列を基準配列として、該基準配列の前記２組のパリティ情報に対応させて、検査対象の生物学的物質の配列について前記２組のパリティ情報を求め、

前記４組のパリティ情報より前記基準配列に対する前記検査対象の生物学的物質の配列の相違部を求めることを特徴とする生物学的物質の配列情報の記録方法。

１１．請求の範囲１に記載の記録方法であって、

前記生物学的物質は、ＤＮＡ、ＲＮＡ、又は遺伝子の少なくとも一部を構成するヌクレオチドであることを特徴とする生物学的物質の配列情報の記録方法。

１２．請求の範囲１に記載の記録方法であって、

前記生物学的物質は、一つのタンパク質の少なくとも一部を構成するアミノ酸であることを特徴とする生物学的物質の配列情報の記録方法。

１３．生物学的物質の配列情報の記録装置であって、

前記生物学的物質の配列情報を読み取る配列読み取り装置と、

前記生物学的物質の配列に対応するテキストデータ、又は該テキストデータを所定の規則に従って変換して得られる数値データを、前記生物学的物質の配列方向に複数行で、かつ前記配列方向に交差する非配列方向に複数列の長さがｍビット（ｍは１６以上の整数）の部分データに分割するデータ配列手段と、

複数行の前記部分データに各行毎に前記非配列方向にガロア体ＧＦ（2^m）上の第１の演算を施して第１組のパリティ情報を求めると共に、複数列の前記部分データに各列毎に前記配列方向にガロア体ＧＦ（2^m）上の第２の演算を施して第２組のパリティ情報を求める演算手段と、

前記第１組及び第２組のパリティ情報を記録媒体に記録する記録手段と

を有することを特徴とする生物学的物質の配列情報の記録装置。

１４．請求の範囲１３に記載の記録装置であって、

前記ガロア体ＧＦ（2^m）上の生成元をαとしたとき、

前記第１組のパリティ情報は、複数行の各行の前記部分データにそれぞれ前記非配列方向に順次α^{sp}，$\alpha^{s(p+1)}$，$\alpha^{s(p+2)}$，…，$\alpha^{s(p+dp)}$（ｓは０以上の整数、ｐは０以上の整数、ｄｐは１以上の整数）を乗算した後、

該演算で得られた複数の積について各行毎に求められた和を含み、

前記第２組のパリティ情報は、複数列の各列の前記部分データにそれぞれ前記配列方向に順次α^{tq}，$\alpha^{t(q+1)}$，$\alpha^{t(q+2)}$，…，$\alpha^{t(q+dq)}$（ｔは０以上の整数、ｑは０以上の整数、ｄｑは１以上の整数）を乗算した後、

該演算で得られた複数の積について各列毎に求められた和を含むことを特徴とする生物学的物質の配列情報の記録装置。

１５．生物学的物質の配列情報を記録したコンピュータ読み取り可能な記録媒体

であって、

前記生物学的物質の配列に対応するテキストデータ、又は該テキストデータを所定の規則に従って変換して得られる数値データを、前記生物学的物質の配列方向に複数行で、かつ前記配列方向に交差する非配列方向に複数列の長さがｍビット（ｍは１６以上の整数）の部分データに分割し、

複数行の前記部分データに各行毎に前記非配列方向にガロア体ＧＦ（2^m）上の第１の演算を施して第１組のパリティ情報を求めると共に、複数列の前記部分データに各列毎に前記配列方向にガロア体ＧＦ（2^m）上の第２の演算を施して第２組のパリティ情報を求め、

前記生物学的物質の配列に関する情報が、前記第１組及び第２組のパリティ情報として記録されたことを特徴とするコンピュータ読み取り可能な記録媒体。

１６．請求の範囲１５に記載の記録媒体であって、

前記生物学的物質の配列に対応する前記テキストデータ、又は該テキストデータに対応する前記数値データの４０ビット以上の長さの数学的な要約値が更に前記記録媒体に記録されたことを特徴とするコンピュータ読み取り可能な記録媒体。

１７．生物学的物質の配列情報の供給方法であって、

前記生物学的物質の配列に対応するテキストデータ、又は該テキストデータを所定の規則に従って変換して得られる数値データを保持する供給者が、前記テキストデータ、又はこれに対応する前記数値データを第１ファイルに記録して保持する第１ステップと、

前記第１ファイルに記録されている前記テキストデータ、又は該テキストデータに対応する前記数値データを、前記生物学的物質の配列方向に複数行で、かつ前記配列方向に交差する非配列方向に複数列の長さがｍビット（ｍは１６以上の整数）の部分データに分割し、

複数行の前記部分データに各行毎に前記非配列方向にガロア体ＧＦ（2^m）上の第１の演算を施して第１組のパリティ情報を求めると共に、複数列の前記部分データに各列毎に前記配列方向にガロア体ＧＦ（2^m）上の第２の演算を施して第２組のパリティ情報を求める第２ステップと、

前記供給者が、前記第１組及び第２組のパリティ情報を第２ファイルに記録し

て保持する第３ステップと、

前記生物学的物質の配列情報のユーザが、通信回線を介して前記供給者より前記第２ファイルに記録されている前記２組のパリティ情報を受け取る第４ステップと

を有することを特徴とする生物学的物質の配列情報の供給方法。

１８．請求の範囲１７に記載の供給方法であって、

前記ユーザが、前記２組のパリティ情報に基づいて検査対象の生物学的物質の配列情報の内の前記供給者の生物学的物質の配列情報との相違部を特定する第５ステップと、

該相違部の配列の復元ができない場合に、前記ユーザが前記通信回線を介して前記供給者より前記第１ファイルに記録されている前記テキストデータ、又は前記数値データの内の前記配列の復元ができない部分の配列情報を受け取る第６ステップと

を有することを特徴とする生物学的物質の配列情報の供給方法。

１９．請求の範囲１８に記載の供給方法であって、

前記ガロア体ＧＦ（2^m）上の生成元をαとしたとき、

前記第１組のパリティ情報は、複数行の各行の前記部分データにそれぞれ前記非配列方向に順次α^{sp}，$\alpha^{s(p+1)}$，$\alpha^{s(p+2)}$，…，$\alpha^{s(p+dp)}$（ｓは０以上の整数、ｐは０以上の整数、ｄｐは１以上の整数）を乗算した後、

該演算で得られた複数の積について各行毎に求められた和を含み、

前記第２組のパリティ情報は、複数列の各列の前記部分データにそれぞれ前記配列方向に順次α^{tq}，$\alpha^{t(q+1)}$，$\alpha^{t(q+2)}$，…，$\alpha^{t(q+dq)}$（ｔは０以上の整数、ｑは０以上の整数、ｄｑは１以上の整数）を乗算した後、

該演算で得られた複数の積について各列毎に求められた和を含むことを特徴とする生物学的物質の配列情報の供給方法。

２０．請求の範囲１８に記載の供給方法であって、

前記供給者は、前記生物学的物質の配列の長さの情報、及び前記配列を表すテキストデータ又は前記数値データの数学的な要約値の情報を前記通信回線を介して閲覧可能な状態にしておき、

前記ユーザは、前記第４ステップの前に前記通信回線を介して前記配列の長さの情報及び前記数学的な要約値の情報を閲覧することを特徴とする生物学的物質の配列情報の供給方法。

２１．生物学的物質の配列情報の記録方法であって、

前記生物学的物質の配列に対応するテキストデータ、又は該テキストデータを所定の規則に従って変換して得られる数値データを、前記生物学的物質の配列方向に複数行で、かつ前記配列方向に交差する非配列方向に複数列の長さがｍビット（ｍは１６以上の整数）の部分データに分割し、

前記部分データの最大値をＮｍａｘ、この最大値Ｎｍａｘよりも大きい素数をＰとして、

複数行の前記部分データに各行毎に前記非配列方向にガロア体ＧＦ（Ｐ）上の第１の演算を施して第１組のパリティ情報を求めると共に、

複数列の前記部分データに各列毎に前記配列方向にガロア体ＧＦ（Ｐ）上の第２の演算を施して第２組のパリティ情報を求め、

前記第１組及び第２組のパリティ情報で前記生物学的物質の配列を表すことを特徴とする生物学的物質の配列情報の記録方法。

２２．請求の範囲２１に記載の記録方法であって、

前記部分データの最大値Ｎｍａｘは（$2^m - 1$）よりも小さい値であり、

前記素数Ｐは、

$2^m > P > Nmax$

の関係を満たすことを特徴とする生物学的物質の配列情報の記録方法。

２３．請求の範囲２２に記載の記録方法であって、

前記ガロア体ＧＦ（Ｐ）上の生成元をδとしたとき、

前記第１組のパリティ情報は、複数行の各行の前記部分データにそれぞれ前記非配列方向に順次δ^{sp}，$\delta^{s(p+1)}$，$\delta^{s(p+2)}$，…，$\delta^{s(p+dp)}$（ｓは０以上の整数、ｐは０以上の整数、ｄｐは１以上の整数）を乗算した後、

該演算で得られた複数の積について各行毎に求められた和を含み、

前記第２組のパリティ情報は、複数列の各列の前記部分データにそれぞれ前記配列方向に順次δ^{tq}，$\delta^{t(q+1)}$，$\delta^{t(q+2)}$，…，$\delta^{t(q+dq)}$（ｔは０以上の整数、

ｑは０以上の整数、ｄｑは１以上の整数）を乗算した後、

該演算で得られた複数の積について各列毎に求められた和を含むことを特徴とする生物学的物質の配列情報の記録方法。

２４．生物学的物質の配列情報の供給方法であって、

前記生物学的物質の配列に対応するテキストデータ、又は該テキストデータを所定の規則に従って変換して得られる数値データを保持する供給者が、前記テキストデータ、又はこれに対応する前記数値データを第１ファイルに記録して保持する第１ステップと、

前記第１ファイルに記録されている前記テキストデータ、又は該テキストデータに対応する前記数値データを、前記生物学的物質の配列方向に複数行で、かつ前記配列方向に交差する非配列方向に複数列の長さがｍビット（ｍは１６以上の整数）の部分データに分割し、

前記部分データの最大値をＮｍａｘ、この最大値Ｎｍａｘよりも大きい素数をＰとして、

複数行の前記部分データに各行毎に前記非配列方向にガロア体ＧＦ（Ｐ）上の第１の演算を施して第１組のパリティ情報を求めると共に、複数列の前記部分データに各列毎に前記配列方向にガロア体ＧＦ（Ｐ）上の第２の演算を施して第２組のパリティ情報を求める第２ステップと、

前記供給者が、前記第１組及び第２組のパリティ情報を第２ファイルに記録して保持する第３ステップと、

前記生物学的物質の配列情報のユーザが、通信回線を介して前記供給者より前記第２ファイルに記録されている前記２組のパリティ情報を受け取る第４ステップと

を有することを特徴とする生物学的物質の配列情報の供給方法。

２５．請求の範囲２４に記載の供給方法であって、

前記ユーザが、前記２組のパリティ情報に基づいて検査対象の生物学的物質の配列情報の内の前記供給者の生物学的物質の配列情報との相違部を特定する第５ステップと、

該相違部の配列の復元ができない場合に、前記ユーザが前記通信回線を介して

前記供給者より前記第１ファイルに記録されている前記テキストデータ、又は前記数値データの内の前記配列の復元ができない部分の配列情報を受け取る第６ステップと

を有することを特徴とする生物学的物質の配列情報の供給方法。

要　約　書

ヌクレオチド又はアミノ酸などの生物学的物質の配列情報を少ないデータ量で近似的に記録する記録方法及び装置である。標準試料ＥのＤＮＡを構成する一列のヌクレオチドの配列を示すテキストデータを所定の変換規則に従ってバイナリーデータに変換し、このバイナリーデータを複数行で複数列のｍビット（ｍ≧１６）の部分データ（Ａ（ｉ，ｊ））に分割する。各行の部分データ（Ａ（ｉ，ｊ））を非配列方向にガロア体ＧＦ（2^m）上で演算して第１組のパリティ（Ｂ１（ｉ）～Ｂ３（ｉ））を求め、各列の部分データ（Ａ（ｉ，ｊ））を配列方向にガロア体ＧＦ（2^m）上で演算して第２組のパリティ（Ｃ１（ｊ）～Ｃ３（ｊ））を求め、これらのパリティ情報によってヌクレオチドの配列を近似的に表す。